BMC Biology

RESEARCH ARTICLE

Open Access

Emergence and influence of sequence bias in evolutionarily malleable, mammalian tandem arrays

Margarita V. Brovkina[1†], Margaret A. Chapman[2†], Matthew L. Holding[3] and E. Josephine Clowney[4,5*]

Abstract

Background The radiation of mammals at the extinction of the dinosaurs produced a plethora of new forms—as diverse as bats, dolphins, and elephants—in only 10–20 million years. Behind the scenes, adaptation to new niches is accompanied by extensive innovation in large families of genes that allow animals to contact the environment, including chemosensors, xenobiotic enzymes, and immune and barrier proteins. Genes in these "outward-looking" families are allelically diverse among humans and exhibit tissue-specific and sometimes stochastic expression.

Results Here, we show that these tandem arrays of outward-looking genes occupy AT-biased isochores and comprise the "tissue-specific" gene class that lack CpG islands in their promoters. Models of mammalian genome evolution have not incorporated the sharply different functions and transcriptional patterns of genes in AT- versus GC-biased regions. To examine the relationship between gene family expansion, sequence content, and allelic diversity, we use population genetic data and comparative analysis. First, we find that AT bias can emerge during evolutionary expansion of gene families in cis. Second, human genes in AT-biased isochores or with GC-poor promoters experience relatively low rates of de novo point mutation today but are enriched for non-synonymous variants. Finally, we find that isochores containing gene clusters exhibit low rates of recombination.

Conclusions Our analyses suggest that tolerance of non-synonymous variation and low recombination are two forces that have produced the depletion of GC bases in outward-facing gene arrays. In turn, high AT content exerts a profound effect on their chromatin organization and transcriptional regulation.

Keywords Chemosensation, Barriers, Xenobiotic metabolism, Genome organization, Isochores, Sequence bias, CpG island, Heterochromatin, Olfaction, Immune system

†Margarita V. Brovkina and Margaret A. Chapman contributed equally to this work.

*Correspondence:
E. Josephine Clowney
jclowney@umich.edu
[1] Graduate Program in Cellular and Molecular Biology, University of Michigan Medical School, Ann Arbor, MI, USA
[2] Neurosciences Graduate Program, University of Michigan Medical School, Ann Arbor, MI, USA
[3] Life Sciences Institute, University of Michigan, Ann Arbor, MI, USA
[4] Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI, USA
[5] Michigan Neuroscience Institute, University of Michigan, Ann Arbor, MI, USA

Background

Reports of newly sequenced genomes frequently describe gene families that have "bloomed," undergoing explosive diversification in the focal species [1–3]. Gene blooms are expansions in cis that result in arrays of dozens or even hundreds of genes. During gametogenesis, these tandem duplication events are thought to arise via incorrect crossovers between paralogues or via non-homologous repair of chromosome breaks [4]. The resulting expansions can confer unique life history traits recognized as definitive characteristics of the species: Examples include Cytochrome p450 genes for plant detoxification in koala and insects, lipocalins for pheromone communication in mouse, NK cell receptors for viral response in bats, keratins for whale baleen, venom production in snakes, and amylase copy number for starch consumption in modern humans [1, 2, 5–11]. The definitive gene family of mammals, the caseins, arose through local duplication of enamel genes [12]. The "birth-and-death" evolution of these gene families also results in high rates of pseudogenization, and some species have lost whole families [13–19]. In the mouse, we have shown that genes in copy-number-variable blooms exhibit extremely high AT content in their promoters and are often located in AT-biased regions of the genome [20].

GC content in mammalian genomes varies markedly at the megabase scale [21]. Since the earliest days of cytology, variation in staining patterns of DNA-binding dyes were apparent across the nucleus (heterochromatin and euchromatin) or along chromosomes (banding patterns). Banding patterns served as the original genetic maps and allowed scientists to link genetic phenotypes to physical positions in DNA [22]. Banding patterns were found to reflect local variation in AT/GC content: Giemsa-staining "G-bands" are AT-biased, and Quinacrine-staining "Q-bands" are GC-biased [22–25]. Early reports suggested that G-bands were depleted for genes; that genes in G-bands tended to be "tissue-specific;" and that genes in Q-bands tended to be "housekeeping genes" [22, 26]. Moreover, human-chimp divergence rates were found to be higher in AT-rich G-bands than in GC-rich Q-bands [27]. In the genome sequencing era, breaks between bands were found to correspond to local transitions in GC content, and bands were found to be composed of smaller "isochores" with locally consistent GC content [28, 29]. While isochore definition has been debated, a representative classification breaks the human genome up into ~3000 isochores of 100 kb–5 Mb that range from 35 to 58% GC [29–32].

The variation in GC content along the chromosome that is observed in mammals is not a general feature of metazoan, animal, or even vertebrate genomes. Both average GC content and the amount of local variation show wide divergence across clades [33, 34], leading to adaptationist speculation that isochore structure serves a function related to endothermy [35]. However, consensus has emerged that GC-biased gene conversion (gBGC) following meiotic recombination, which occurs in most eukaryotes, is one important contributor to isochore emergence in mammals [33, 36]. In this process, crossovers are statistically more likely to result in gene conversion towards more GC-rich sequences, resulting in a higher likelihood of inheriting higher-GC alleles. As stated by Pouyet and colleagues, "The gBGC model predicts that the GC content of a given genomic segment should reflect its average long-term recombination rate over tens of million years" [37, 38]. In this model, the isochores themselves do not serve an adaptive function, but rather have emerged due to the molecular genetic ("neutral") forces of meiosis. Over evolutionary time, the GC-increasing effect of recombination is counteracted by the AT-increasing effect of point mutation due to mutation of fragile cytosines to thymines [39–44]. As the rate of recombination and cytosine loss may themselves be influenced by sequence context (i.e., recombination more likely in GC-rich regions, cytosine mutation more common in AT-rich regions), positive feedback could have caused large genomic regions to diverge [27, 37, 39, 45]. However, the influence of these neutral forces on the genes contained within AT- versus GC-biased isochores has not been described.

Here, we examine the local and isochore-level AT/GC content of human protein-coding genes. Genes located in AT-rich regions of the genome have unique and consistent characteristics: They are copy-number-variable families located in tandem arrays, are expressed in terminally differentiated cells, are cell surface or secreted proteins, lack CpG islands in their promoters, and often have stochastic or variegated expression. These protein families are overwhelmingly involved in the "input-output" functions of an organism: sensation of the environment, protection from the environment, consumption of the environment, and production of bodily fluids. AT- versus GC-skewed isochores differ in their patterns of histone marks; likely as a consequence, they differ in their replication timing, associate in nuclear space with other isochores of the same type, and occupy different domains within the nucleus [23, 46–48]. The distinct treatment of AT- versus GC-rich isochores by the molecular machinery of the mammalian cell means that the genes located in AT-rich isochores must experience distinct molecular events from those located in GC-rich isochores.

Next, we ask how mammalian genes with outward-looking functions came to be located in AT-rich regions of the genome. By comparing gene blooms of different sizes within the human genome and across mammalian species, we find that AT content is not necessarily inherited from the ancestral species but can emerge with cluster expansion. Many of the mutagenic and repair processes that contribute to the neutral mutation spectrum are likely to differ in strength across AT-rich and GC-rich regions (reviewed in [49]). Using human population genetic data, we analyze allelic variation, patterns of point mutation, and recombination in human genes located in AT- versus GC-biased isochores. We find that genes in paralogous clusters are subject to less recombination than genes located near non-paralogues. Recombination may be dangerous in gene clusters due to the potential for chromosome rearrangements or within-cluster ectopic exchange that leads to duplications or deletions. It could also separate genes in large families from locus control regions they depend on for expression.

In addition, we find that while genes in AT-biased isochores have high sequence diversity among humans and divergence across species, they do not currently exhibit excess de novo point mutations (DNMs); as expected from the neutral mutation spectrum, GC-rich isochores and their genes exhibit higher DNM rates. Instead, genes in AT-biased isochores appear to have accumulated sequence variants over evolutionary time. The outward-looking genes in these allelically variable tandem arrays lack CpG island promoters and drive the known “tissue specificity” of genes that lack islands [26, 50, 51]. We hypothesize that the functional roles of genes whose protein products interface with unpredictable and rapidly changing molecules in the environment (“outward-looking genes”) make them particularly likely to tolerate (or benefit from) non-synonymous variation. Diminished purifying selection on these genes is expected to shift GC content down over evolutionary time due to deamination of cytosine leading to C->T transitions [49]; intolerance of recombination would prevent gBGC from shifting GC content back up [44]. We propose a model in which reduced recombination and diminished selection on point mutations act together to strand arrays of outward-looking paralogues in wells of low GC content (Fig. 1). Loss of CpG islands and residence in AT-rich genomic regions predisposes these genes to exotic forms of highly tissue-specific transcriptional regulation [52–56].

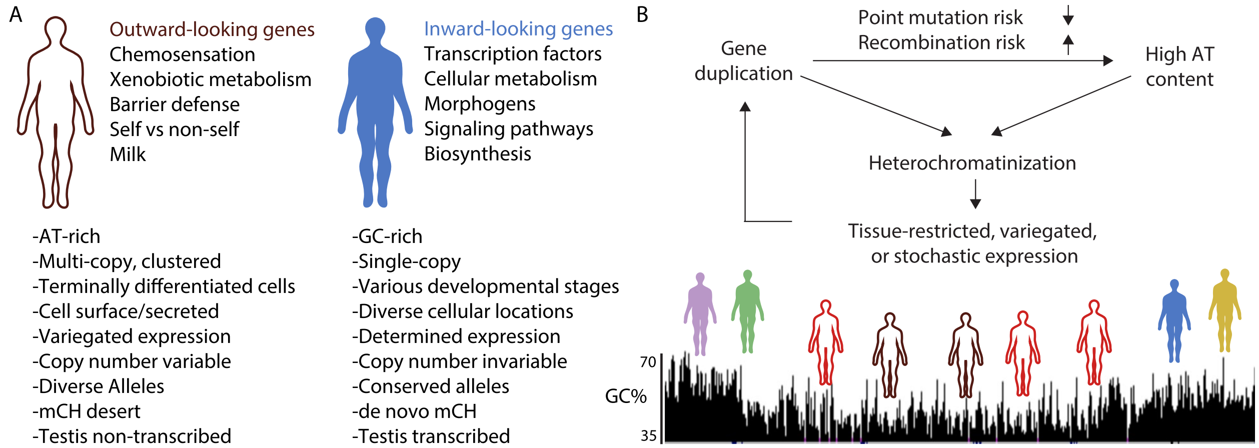

Fig. 1 Outward- versus inward-looking genes. **A** Summary description of the characteristcs of inward-looking and outward-looking gene families and their genomic distinctions. This table is inspired by Holmquist [57]. Descriptions derive from our analyses here and from references cited throughout the text. **B** Model of possible relationships between expansion of gene families in cis, genomic architecture, selective forces, and mode of expression. We hypothesize that as gene families duplicate in cis, selection on the amino acid sequence of individual family members weakens, while selection against recombination strengthens. Together, these effects would result in loss of GC bases over evolutionary time. Once a tandem array is AT-rich, it is more likely to be heterochromatinized and acquires highly tissue-specific expression patterns. The “quarantining” of expression reduces the phenotypic consequences of change in copy number, allowing further rounds of gene gain; nevertheless, the numbers of intact genes in certain families correlates with natural history and is likely under selection [58–61]. This mode of “birth-and-death evolution” also results in frequent gene loss through pseudogenization [17, 18, 62]. We expect that while forces described here are important contributors to emergence of sequence bias, our model is incomplete—there are likely to be additional neutral or selective mechanisms that make important contributions to the emergence of sequence bias in outward-looking tandem arrays. These could include the basal sequence content of a gene prior to any duplication, sequence effects of the molecular mechanism that produces gene duplication, differential amino acid usage in different kinds of proteins, or selective mechanisms that preferentially retain duplicates with weak promoters. Ultimately, these forces result in the observed gene content in mammalian isochores, where outward-looked arrays (reds and browns) are enriched in AT-rich regions, and single-copy genes with inward-looking functions (other colors) in GC-rich isochores

Results

Characterizing a set of human isochores and their gene contents

We first comprehensively characterized the types of genes located in AT- versus GC-biased isochores and assessed if genes in AT-biased isochores are more likely to have numerous paralogues as neighbors. Isochore boundaries have been computed for previous genome assemblies, including for hg38 [29–31, 63]. While previous methods of isochore annotation binned the genome into 100 kb pieces or used manual inspection to annotate isochore ends, we used a segmentation algorithm which detects transitions in GC content and created a UCSC Genome Browser track for easy visualization of isochore assignments versus other genomic elements [64]. Cozzi et al. compare various methods of isochore assignment, including GC-Profile, showing that differences are subtle and are most prevalent in the mid-ranges of GC content. We tested three resolutions and selected a map that visually matched 100 kb–Mb transitions in GC content using the %GC track on the UCSC Genome Browser (Fig. 2A, B). At this resolution, we call 4328 isochores; these range from ~30–70% GC and most are between 100 kb and 5 Mb (Fig. 2C, D, Additional files 1 and 2, Additional file 3: Fig. S2A-C).

We next defined the home isochore of each gene in the NCBI MANE set. The MANE set includes one promoter and splice isoform for each intact, protein-coding gene and omits pseudogenes and complex "gene parts" such as V, D, and J segments of the B and T cell receptors [67]. On average, higher-GC isochores were more likely to contain protein-coding genes and had higher gene density (Fig. 2C, Additional file 3: Fig. S2A-B) [22]. To compare features across isochores, we ordered isochores by GC% and divided them into ten groups (deciles) of ~400. To test whether tandemly arrayed "gene blooms" were associated with AT-rich isochores, we used Shannon's H to measure gene name prefix diversity in isochores with at least ten genes (Fig. 2D). Gene names were least diverse in AT-rich isochores, consistent with the presence of tandem arrays in these isochores. We extended our analysis to isochores with fewer genes (at least five) and again found that gene arrays were less common in high-GC isochores (Fig. 2E, Additional file 3: Fig. S2H-I). While AT-rich isochores are longer, gene diversity is lower in AT-rich isochores across the length distribution and for isochores with different numbers of genes (Additional file 3: Fig. S2F-G). Arrays in AT-rich isochores contained more paralogues, and the kinds of proteins present in AT-rich arrays versus GC-rich arrays were different (Fig. 2E, Additional file 3: Fig. S2F): Arrays in AT-rich isochores often served sensory, digestive, or barrier functions, while arrays in higher-GC isochores most often contained *HOX* and *ZNF* transcription factors or arrays of histones (Fig. 2E, Additional file 3: Fig. S2I).

Only 25% of genes are located in isochores <40% GC. What sorts of gene families have bloomed in these isochores? As GO analysis is biased by the annotations that are available, especially for extremely tissue-specific genes, we simply searched for common gene name prefixes in isochores <40% GC (Fig. 2F). Gene families with at least four members in high-AT isochores were overwhelmingly involved in chemosensation (*OR*, *TAAR* , *TAS2R*, *MS4*); xenobiotic metabolism (e.g., *AMY1*, *UGT2*, *ADH*); and immune, defense, and barrier functions (e.g., *KLR*, *IFN*, *DSC/DSG*). Human ORs were also shown previously to be located in AT-biased isochores [68]. Examples of high-GC isochores with diverse gene members and high-AT isochores with repetitive gene members are shown in Fig. 2A, B and Additional file 3: Fig. S2D-E. While immunoglobulin parts do not appear in the MANE gene set, arrays of immunoglobulin V regions are also highly AT-biased (Additional file 3: Fig. S2G). These analyses demonstrate that AT-biased regions of the human genome contain tandem arrays of genes with outward-looking functions. We invite readers to explore these patterns further on our TrackHub, listed under "Availability of data and materials".

(See figure on next page.)

Fig. 2 AT-rich isochores in the human genome contain tandem arrays of genes with outward-looking functions. **A**, **B** UCSC Genome Browser screenshots showing GC% trajectory [65], CpG islands [66], our isochore calls, and simplified gene models. Gene models are colored according to *k*-means clusters described below. Gene name prefix diversity (Shannon's H) is shown for the highlighted isochores. **C** Relationship between isochore GC% and gene contents. Isochores shown in **A**, **B** are highlighted. Colors show how we binned isochores into deciles according to GC%. Dashed line shows the mean GC content of the human genome (41%). **D** Boxplots representing gene name diversity (Shannon's H) of gene-rich isochores across GC% deciles shown in **C**. Red points indicate the mean prefix diversity of the decile; black lines show medians. Here and throughout, most groups are statistically different from one another, except for adjoining groups; full statistical comparisons are presented in Additional file 4. **E** Comparison of number of genes in an isochore versus number of different gene name prefixes represented in that isochore, for isochores of at least five genes. Each dot is an isochore; isochores falling below the trend line contain clusters of genes with the same prefix. Isochores are labeled by the most common gene name prefix in that isochore. **F** Gene families with at least four members in isochores < 40% GC. Twenty-five percent of all genes are in isochores less than 40% GC (black line). Red bars depict proportions of genes with that prefix that are located in < 40% GC isochores. Gene family prefixes shown in black text are enriched in AT-rich isochores, while those shown in gray text (e.g., *KRTAP*, *TRIM*, *SCN*) have multiple family members in AT-rich isochores but are not enriched there. Functions of these gene families are marked at left, and total number of genes with that prefix in the MANE set are shown at right

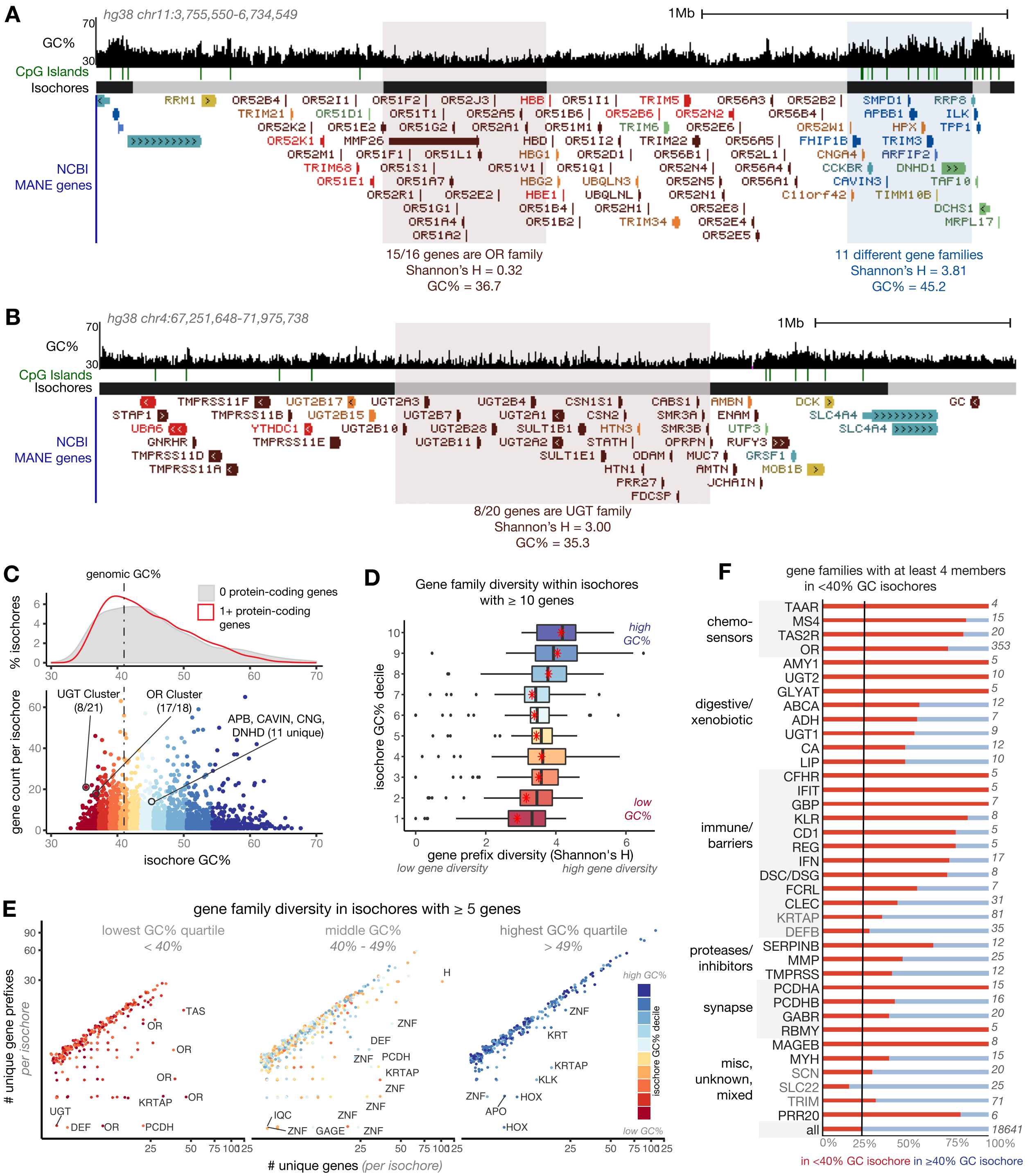

Fig. 2 (See legend on previous page.)

Categorizing human genes according to local patterns of AT/GC content

We show above that tandemly arrayed genes serving outward-looking functions are enriched in AT-rich regions of the human genome, as they are in mouse [20]. We sought next to ask whether the regulatory and transcribed regions of genes, which occupy a fraction of genome space, also vary in GC content across different kinds of genes, and whether GC content of local gene features follows that of the isochore context. The null expectation of this analysis is different from a statistical genetic versus a molecular point of view. From a statistical standpoint, the expectation would be that GC% would co-vary between gene parts (e.g., flanking regions, promoter, coding sequence) and their isochore context. From a molecular point of view, transcribed units and regulatory regions would be expected to have a GC content aligned with their regulatory or protein-coding function and would not necessarily match their genomic location. We calculated GC% in 50-bp sliding windows along the transcriptional unit (transcription start site to transcription end site, TSS-TES) and 1-kb flanking regions for genes in the MANE set [65, 67]. Our analysis here includes introns, but clustering on exons and flanking regions produced similar results (Additional file 3: Fig. S3A-B). We used iterative *k*-means clustering to group the 18,640 MANE genes into 3×3 sets (Fig. 3A, Additional file 3: Fig. S3A). The top-level clusters (1-, 2-, 3-) reflect overall differences in AT content in different genes (Additional file 3: Fig. S3A), while the subclusters (1.1, 1.2, 1.3, etc., Fig. 3A) reflect variation in AT content of the promoter, transcriptional unit, and 3′ region. To capture both the broad isochore context of genes and their local sequence features, we use both the isochore AT/GC metric and the local sequence-based *k*-means clustering throughout this study; each gene in the MANE set is assigned uniquely to one home isochore and one *k*-means cluster (isochore deciles 1–10, red-blue palette; *k*-means clusters 1.1–3.3, rainbow palette). Cluster and isochore assignments and other gene-linked data are provided in Additional files 5 and 6.

The power of this approach is that it captures patterns of feature GC% in relation to one another: For most gene categories, a sharp rise in GC content marks the approach of the TSS, while the transcribed region and the region 3′ of the TES share lower GC content. This GC rise at the TSS clearly corresponds to the promoter. In this context, the paltry GC enrichment at the promoters of genes in cluster 3.3 (and to a lesser extent 3.1) is extremely stark (Fig. 3A).

Based on high-confidence annotation of transcription start sites, we showed previously that mouse olfactory receptor promoters share this GC-poor pattern [20]. At that time, the TSS's of other highly tissue-specific genes had not been mapped. Current human annotations in the MANE set are high-confidence, curated gene models. Nevertheless, many of the genes in cluster 3.3 are extremely tissue-specific (see below) and have less supporting mRNA data than more widely expressed genes. We identified 240 genes in the MANE set (1.3%) that lack annotated 5′ UTRs, i.e., where the annotated transcription start site and translation start site are the same. While these were indeed mostly contained in cluster 3.3 (150 of 1636 genes in 3.3, 9%), there was no difference in promoter GC content between these suspect gene models and other genes in cluster 3.3 (data not shown). Indeed, our manual inspection of available RNA data for a subset of these genes suggests that the transcription start sites are correct, but that translation likely initiates at a downstream ATG.

We next examined how a gene's isochore GC context relates to the sequence content of its promoter and coding region. We found that patterns of local GC content of genes predicted the GC content of their home isochore, consistent with the statistical genetic null hypothesis, but very surprising considering the functional implications (Fig. 3B). The GC content of a gene with its flanking regions (gene extent with 25 kb on each side) correlated closely with the GC content of its whole isochore (Additional file 3: Fig. S3C). Individually, promoter and coding sequence GC% were also positively correlated with isochore sequence content, but the correlation coefficients were weaker: A subset of genes in AT-rich isochores have GC-rich promoters or GC-rich coding sequences (Additional file 3: Fig. S3D-E).

Genes in clusters 3.1 and 3.3, lacking GC enrichment in their promoters, were highly enriched for the same functional categories as were genes in AT-rich isochores: chemosensation, xenobiosis, and defense/barriers. Indeed, as can be seen in Fig. 2A, B and Additional file 3: Fig. S2D, sometimes entire arrays were members of cluster 3.3 (brown color). To systematically test this, we plotted the promoter GC content distribution of genes in categories we term "outward-looking" (chemosensation, defense, xenobiosis, barriers) versus "inward-looking" (e.g., transcription, kinase function, morphogens). Outward-looking genes have AT-rich promoters while inward-looking genes have GC-rich or average promoters (Fig. 3C). We manually annotated common prefixes and enrichment of genes in cluster 3.3 (Fig. 3D): This group included all the chemosensory families, many sets of digestive and detoxifying enzymes, and several receptor arrays in the immune system and skin. It also included clustered protocadherins, which share transcriptional regulation patterns with chemosensors. In accordance with the preponderance of tandemly arrayed genes found in cluster 3.3, we found that genes in this cluster were housed in fewer unique isochores and had lower name diversity than those in the other *k*-means clusters (Fig. 3E, Additional file 3: Fig. S3C).

Finally, we asked whether being located near paralogues could predict local sequence features (Fig. 3F). Indeed, genes near 1–4 neighbors with the same prefix had more AT-rich promoters than genes not located near paralogous genes, and genes with more than four same-prefix neighbors had AT-elevated promoters, coding regions, and flanking regions compared to both genes in small clusters and singletons. The striking coordination of isochore, promoter, and coding sequence AT content in tandemly arrayed "outward-looking" gene families prompted us to investigate how this pattern relates to the evolution of tandem arrays.

Increasing AT content during evolutionary expansion of tandem arrays

As shown above, we find that both the regional and local AT content is high in tandemly arrayed gene clusters in human. AT bias could have emerged as gene families expanded or could have been pre-existing and supported molecular mechanisms of gene duplication. To ask whether AT content rises with the number of paralogous genes in a tandem array, we sought to track copy-number-variable tandem arrays across mammals. Assessing the evolution of tandem arrays is difficult, as large proportions of local paralogues can be species-specific duplications (Additional file 3: Fig. S4C and [62, 69–71]). Therefore, instead of seeking to identify true homologous genes across different mammalian species, we used synteny analysis to follow whole arrays over evolutionary time. We searched for large gene arrays in human that could be identified across diverse mammals using microsynteny, where conserved heterologous genes serve as "bookends" bounding the ends of the paralogous tandem array.

While copy-number-invariable gene arrays, such as the HOX clusters, were easy to track across all mammals, as expected, the copy-number-variable gene arrays that would be informative for this analysis showed frequent assembly errors, micro-inversions, and array invasions by other heterologous genes. For example, while we observed that the expanded KLR array of NK cell receptors in the fruit bat *Rousettus aegyptiacus* and pheromone-associated MUP array in *Mus musculus* each had sharply higher AT content than surrounding genomic regions, each array had a large assembly gap [9, 72]. While this limited the number (and, indeed, the extremity) of arrays that we were able to track, we identified six copy-number-variable arrays that were suitable for analysis: an OR array adjacent to the hemoglobin beta genes [73]; the Cyp2ABGFST cluster of xenobiotic enzymes; the SERPINA cluster of defensive protease inhibitors; the "epidermal differentiation complex" (EDC), containing skin proteins in the LCE, S100, and SPRR families; and two clusters of keratin-associated proteins (KRTAPs) which form epidermal appendages such as hair, nails, and claws. These six clusters have diverse median GC content as can be seen in Fig. 4A–H. We also included the copy-number-invariable HOXA cluster for comparison.

For each of these clusters, we defined bookend genes as indicators of synteny that must be present for a particular species to be included in analysis. In mammals for which we could find these bookends on the same scaffold or chromosome, we computed the number of genes between bookends, spot-checking species whose counts were markedly higher than those of other species. The number of paralogues in each cluster is plotted across the mammalian clade in Additional file 3: Fig. S4C, showing the frequency of lineage-specific cluster expansions and contractions. Our full annotations of the contents of these clusters are provided in Additional file 7.

For each cluster in each species, we calculated the GC% between the bookend genes; relationships between GC% and gene count appeared to be log-linear. We performed phylogenetic least squares (PGLS) analysis to test for a relationship between GC content and log(paralog number) while controlling for phylogenetic relatedness in the multi-species dataset (Fig. 4A–H). For the OR, EDC, and SERPINA clusters, AT content rose significantly with paralogue number (PGLS $P < 0.05$), suggesting that AT content can rise as the number of paralogs in tandem arrays grows. Certain arrays in certain species will have latent assembly or annotation errors; however, for the HBB-OR array, we manually counted a subset of species

(See figure on next page.)

Fig. 3 Genes with outward-looking functions have high local AT content. **A** GC content trajectory for human protein-coding genes in the MANE set. Genes were subdivided by iterative *k*-means clustering. At top, the average GC content trajectory for each *k*-means cluster is shown as a line graph. At bottom, each gene is a row and GC content across the transcriptional unit and flanking regions is depicted from red (high AT) to blue (high GC). Rainbow colors assigned to each *k*-means cluster here will be used throughout. **B** Relationship between *k*-means cluster assignment and home isochore GC% for each gene in the MANE set. Red lines depict medians. **C** GO term distribution by promoter GC content for genes in the MANE set. Genes with immune, barrier, chemosensory, and xenobiotic functions have AT-skewed promoters. Genes with developmental and intracellular functions have GC-skewed promoters. Gray shading shows promoter GC content distribution of the whole MANE set. **D** Gene name prefixes enriched in cluster 3.3. Families shown have at least four members in cluster 3.3; proportion of the family located in this cluster is depicted in brown bars. Less than 10% of genes are in cluster 3.3 ("all"). **E** Gene prefix diversity (Shannon's H) is lowest in cluster 3.3. **F** Distribution of promoter, intron, flanking sequence, and coding sequence GC% across genes which do not have paralogous neighbors in their home isochore (gray, genes with 0 tandem neighbors), genes which exist in small local duplications (red, genes with 1–4 tandem neighbors), and genes which exist in large multigene arrays (brown, genes with 5 or more tandem neighbors). A subset of genes without introns appears as 0's

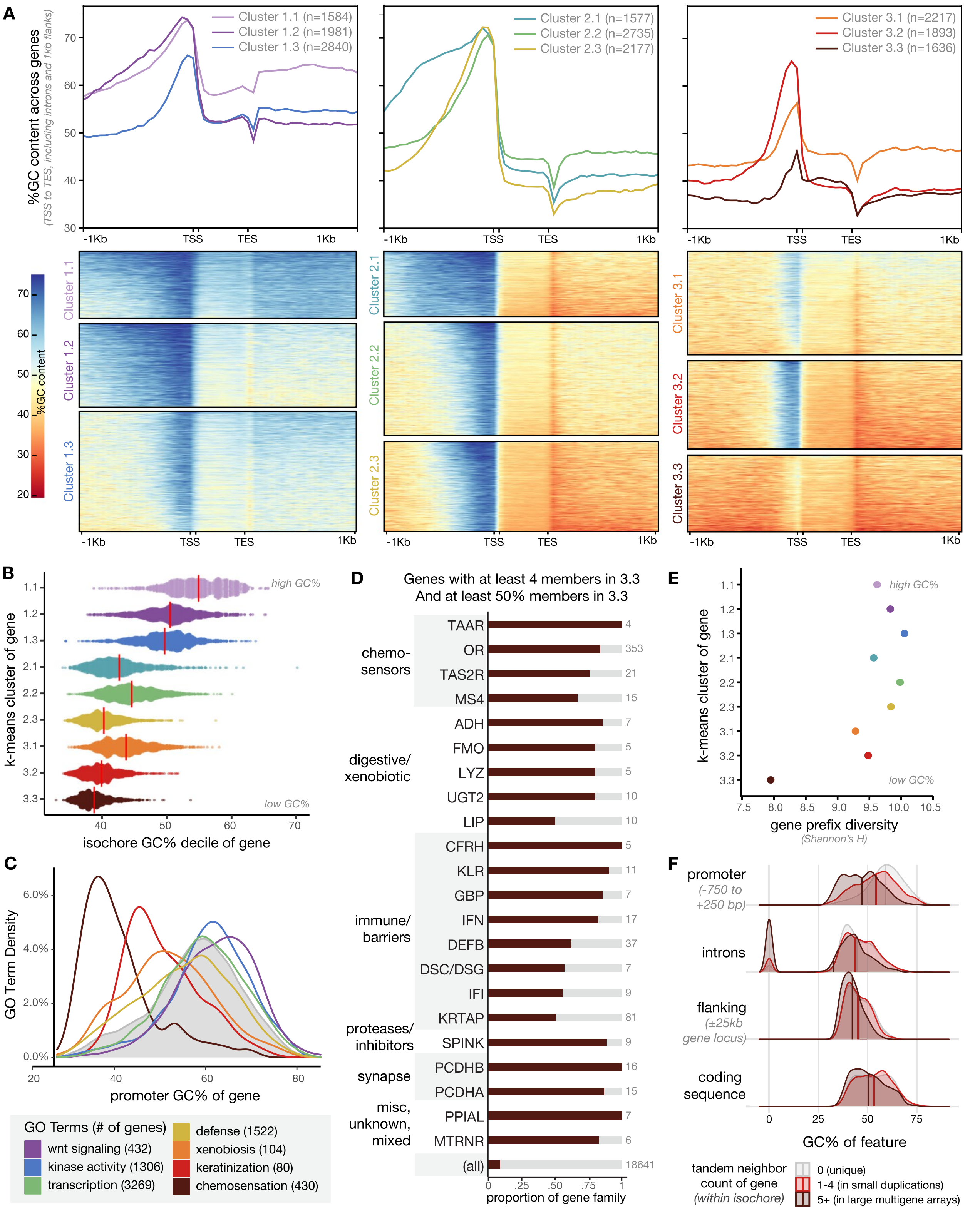

Fig. 3 (See legend on previous page.)

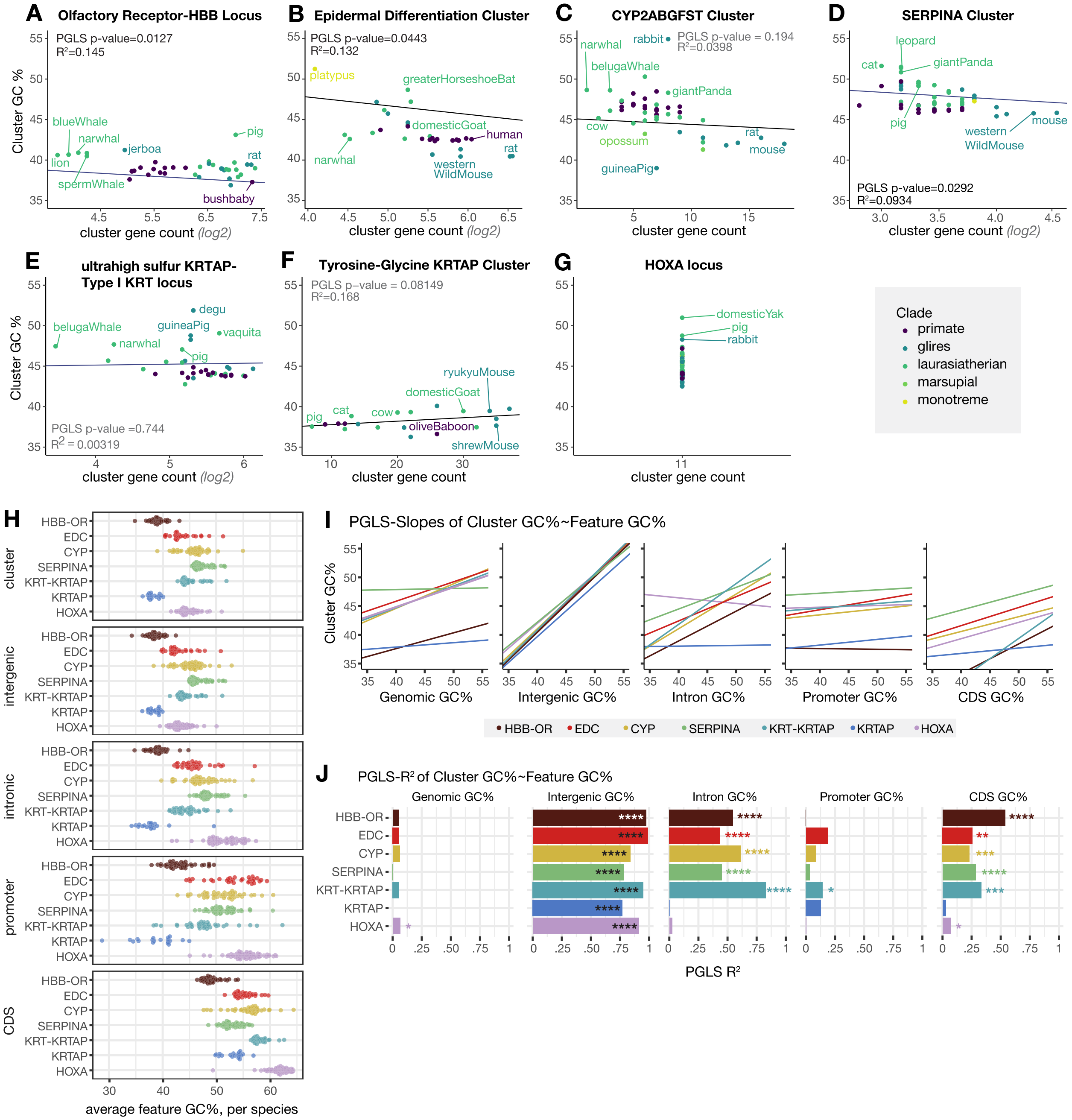

Fig. 4 AT content is correlated with tandem array expansion. **A**–**G** Comparison of number of **A** OR genes in the Hemoglobin β (*HBB*) cluster, **B** LCE, SP100, and SRR genes in the Epidermal Differentiation cluster, **C** CYP2A genes in the CYP2ABGFST cluster, **D** SERPINA genes, **E** KRTAPs and KRTs in the Type I KRT locus, **F** KRTAPs in the high tyrosine-glycine KRTAP cluster, and **G** the HOXA locus with cluster GC% across mammals. All *p*-values report results of phylogenetic least squares analysis (PGLS). **H** Average GC% across the cluster, intergenic, intronic, promoter, and coding sequence of each tandem array. Each point represents one mammal species from panels **A**–**G**. **I** PGLS slope values and **J** R-squared values of phylogeny-corrected correlation of feature GC% to cluster GC%. Asterisks represent magnitude of PGLS *p*-values for each relationship, i.e., one asterisks represents a *p*-value between 0.05 and 0.01, two represent a *p*-value between 0.01 and 0.001

in parallel and saw the same trend (Additional file 3: Fig. S4A). The HOXA cluster showed variation in GC% despite maintaining the same number of paralogs in the cluster over evolutionary time; this variation appears to be predicted by evolutionary variation in the length of the cluster (i.e., number of base pairs between bookends, not shown). We were able to follow the HBB-OR cluster to non-mammalian amniotes and found a sharp rise in

GC content as the number of ORs in the region melted to 0 (Additional file 3: Fig. S4A-B).

The raw GC% of each cluster was poorly correlated with variation in genome-wide GC content across species (Fig. 4I, J). However, each cluster type maintained a consistent relationship to the overall GC% of the genome: Across diverse mammalian species, the OR cluster and KRTAP cluster were almost always more AT-rich than the genome as a whole, while the other five clusters were almost always more GC-rich than the genome as a whole (not shown).

What portions of these genomic regions drive the variation in local GC content over evolutionary time? To assess this, we developed a series of scripts we call TandemClipR (see "Availability of data and materials"), which divided each cluster in each species into CDS, intron, promoter, and intergenic regions. To avoid the "false-TSS" problem described above, we included only genes with annotated 5′ UTRs in our promoter analyses. We found that the GC% of each sub-region was positively correlated with the overall cluster GC% (Fig. 4H–J). As we described above, this is expected from a statistical point of view, but surprising from a molecular point of view given that coding regions and promoters perform important molecular work. For example, only some amino acids can be coded with high-GC codons. Variation in intergenic regions contributed the most to cluster GC%, as expected based on their comprising a preponderance of the sequence. Despite varying with cluster GC%, coding regions, introns, and promoters had consistently higher GC content than the cluster as a whole, consistent with their functional constraints (Fig. 4H).

Finally, we asked how the GC% of sub-regions of the cluster were related to the number of paralogues in the cluster (Additional file 3: Fig. S4D-E). For the clusters whose AT content rose with local paralogue number, the AT content of each substituent portion of the sequence (promoters, coding regions, introns, intergenic regions) also rose. The particular feature that correlated best with paralogue number varied across gene families, and no individual feature was consistently more correlated with paralog number than GC% of the whole cluster. In sum, we find that cluster AT content can rise as paralogous clusters bloom over evolutionary time; remarkably, these trends affect all the sequence components of the cluster, suggesting neutral or selective mechanisms that act cluster-wide. We hypothesize that local gene contents, particularly paralogous clusters of genes, can influence the emergence of isochores differing in GC content. Next, we provide a model for how this could have come about.

Accumulation of Coding Sequence Diversity

As we show above, many types of outward-facing genes have extremely divergent paralogue number across mammalian species (Additional file 3: Fig. S4C). Anecdotally, genes in these families also exhibit extreme allelic diversity and copy number variations among humans, and polymorphisms in these genes underlie human phenotypic variation in drug metabolism, sensory perception, and immune response [1, 11, 62, 74–79]. Outward-looking tandem arrays evolve by birth-and-death evolution and are littered with pseudogenes [17, 18, 62]. A variety of segregating loss-of-function polymorphisms in these gene families have been described in humans [13–16]. Colloquially, outward-looking genes are so diverse in copy number and sequence that a first step in GWAS is often to "throw out the ORs".

Previous reports, including ours, have speculated that partitioning inward- and outward-looking genes into different parts of the genome could enable a higher ongoing mutation rate in AT-rich, outward-looking genes; however, point mutation rate is in general positively correlated with GC content because cytosines are especially mutation-prone [20, 49, 80, 81]. In the first model, AT content would facilitate mutagenesis of genes in these families. In the second model, relatively high point mutation drift in large gene families could have contributed to their shift to higher AT content over evolutionary time.

To test between these possibilities, we first systematically examined the degree of coding sequence variation in human genes grouped by AT/GC content or by degree of local tandem gene duplication, using the gnomAD dataset of rare single-nucleotide variants ascertained from whole exome sequencing of >100,000 unrelated people (gnomAD v2.1.1, Fig. 5A, B) [82]. Genes in AT-rich isochores and *k*-means cluster 3.3 are highly enriched for non-synonymous versus synonymous variants (Fig. 5A, B). Genes near paralogous neighbors were also enriched for non-synonymous variants (Fig. 5E, F). We note that use of rare variants profoundly understates the allelic variety in outward-looking genes, which exhibit radical common variation and high rates of copy number and structural variation. For example, any two humans are estimated to have function-changing variation (loss of function, change in expression level, or change in tuning) in 30% of their olfactory receptor genes [83, 84].

Does the high polymorphism of genes in outward-looking tandem arrays result from differential mutation or selection versus inward-looking genes? We first examined synonymous variants from gnomAD as a proxy for mutations. We see that AT-rich genes have *fewer* synonymous variants across unrelated people than do GC-rich genes (Fig. 5C, D, Additional file 3: S5A-D). This is consistent with point mutation rate being grossly driven

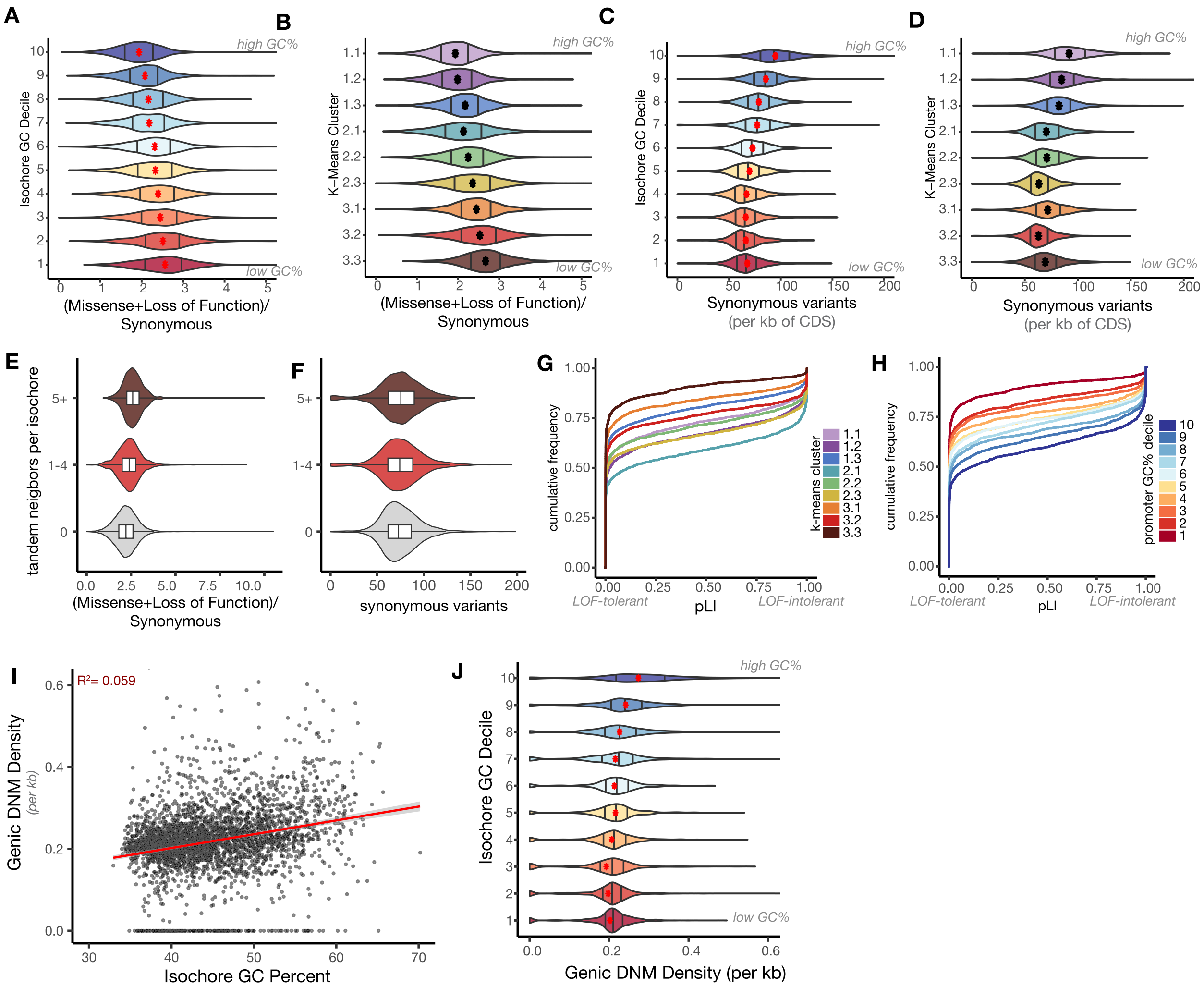

Fig. 5 AT-rich genes have high diversity despite experiencing moderate mutation rates in the present. **A**, **B** Ratio of non-synonymous (missense plus loss of function) versus synonymous rare variants in the MANE gene set identified in gnomAD v2.1.1 exome sequencing of > 100,000 unrelated individuals [82]. Genes are binned by isochore decile (**A**) or *k*-means cluster (**B**), and dots indicate means. gnomAD rare variants are defined by < 0.1% allele frequency. **C**, **D** Raw counts of rare synonymous variants per gene in gnomAD v2.1.1 binned by isochore decile (**C**) or *k*-means cluster (**D**). **E**, **F** Variant rates from gnomAD v2.1.1 for genes with or without paralogous neighbors. Box plots show median and mid-quartile distribution. **G**, **H** Cumulative frequency distribution plots of gnomAD pLI (likelihood that a gene is loss-of-function intolerant) relative to a gene's *k*-means cluster assignment (**G**) or promoter GC% (**H**) [82]. **I**, **J** Number of de novo point mutations observed per kb across the genes (TSS to TES) within an isochore relative to isochore GC% (**I**) and isochore GC% binned by decile (**J**). ~700,000 DNM calls are pooled from all ~11,000 trios sequenced to date [85]

by deamination of cytosine, especially in the C^{me}G context: AT-biased genes have essentially run low on CpG dinucleotides to mutate [39–43, 86, 87]. Incidentally, we observe that olfactory receptor genes have higher rates of variant calls than do other AT-rich genes (Additional file 3: Fig. S5F). This inflated rate could reflect an unknown mutagenic process but could also result from incomplete knowledge of the full human "OR-ome" and incorrect assignment of homology relationships.

The enrichment of non-synonymous variation in outward-looking genes relative to low levels of synonymous variation could result from different selective processes acting on single- versus multicopy genes. Single-copy genes are uniquely responsible for a particular biological process, while tandem arrays may distribute a particular process over a large set of family members (i.e., subfunctionalization). Deleterious mutations to single members of such large gene families are therefore more likely to be of scant phenotypic consequence. Indeed, we find that genes with more tandem neighbors have higher rates of non-synonymous versus synonymous variants (Fig. 5E, F, Additional file 3: Fig. S5E). To test how local and isochore

GC content relate to the functional dispensability of genes, we used the gnomAD metric of "loss-of-function intolerance" (pLI): genes have high pLI if deleterious mutations are depleted from the population [82, 88]. Genes predicted to be loss-of-function intolerant were enriched in GC-rich isochores, had GC-rich promoters, and were absent from cluster 3.3 (Fig. 5G, H, Additional file 3: Fig. S5G); AT-skewed, outward-looking genes had lower probabilities of being loss-of-function intolerant, i.e., they are relatively dispensable. We emphasize that genes in all categories still experience some level of purifying selection, as the observed rate of loss of function alleles is still less than expected from the null mutation spectrum (Additional file 3: Fig. S5F). For example, the total number of intact members of various chemoreceptor subfamilies has been observed to correlate with ecosystem or diet [58–61, 89].

To examine modern mutagenesis patterns directly, we sought to measure the rate of de novo mutations that occur in genes in AT- versus GC-biased regions of the human genome. Whole genome sequencing of two parents and a child (trios) enables detection of de novo mutations (DNMs). We used a recent dataset that compiles DNMs from >11,000 trios, nearly all those who have been sequenced to date [85]. While these ~700,000 total DNMs remain sparse relative to the size of the genome, they approximate a record of mutagenesis that has yet to be operated on by external selection. Pooling DNMs across each isochore and across transcriptional units (TSS to TES) within that isochore, we found that both genic and isochore-wide DNMs were more common in higher-GC isochores (Fig. 5I, J, Additional file 3: Fig. S5H-I). This is consistent with sequence-based predictions, prior findings, and our analysis of gnomAD synonymous variants within genes [90, 91]. We note that while both DNM and synonymous variant rates are highest in high-GC isochores, there is less variation across deciles 1–5. This may suggest that late replication timing, lack of transcription-coupled repair, or higher rate of methylation of CpG sites in AT-rich regions somewhat counterbalance CpG prevalence in GC-rich regions.

Regions that are extremely AT- or GC-rich are sequenced somewhat less often by short-read methods than are sequences of middling GC content. Therefore, higher read depth is required to saturate detection of DNMs in sequence-biased regions [90]. To assess the saturation and consistency of DNM calls across isochores of varying GC content, we separated the aggregate DNM dataset by original study and repeated our analyses (Additional file 3: Fig. S5J) [90, 92–96]. We found that the trend of more DNMs in higher-GC isochores replicated in most individual studies and that this trend became stronger in studies with higher sequencing depth, as expected. In addition, individuals with certain disorders or diseases, such as autism spectrum disorder, are enriched among the individuals who have been assessed for DNMs; these differences did not drive the variation in DNM rate across isochore GC categories (data not shown) [85].

Together, comparison of de novo mutations in meiosis and single-nucleotide variants in unrelated humans both support the conclusion that despite their high divergence and diversity, AT-biased genes experience fewer contemporary mutations overall than do GC-biased genes. It is therefore unlikely that genes are quarantined in AT-rich regions of the genome to facilitate higher mutation rates. Instead, diminished purifying selection on members of outward-looking gene blooms could underlie both diversity and increased AT content.

Recombination and PRDM9 Binding

Recombination may be the primary influence on large-scale patterns of AT/GC content due to GC-biased gene conversion, which elevates GC content. AT-biased chromosomal bands are described as experiencing less recombination than GC-biased bands [27, 33, 37]. We sought to test how recombination rates relate to GC content of inward- versus outward-looking genes. We therefore used whole genome sequencing data from human trios to measure recombination in isochores of different GC contents and in their constituent genes [92].

Crossovers appeared rare within gene blooms (Fig. 6A, Additional file 3: Fig. S6B). We calculated a relative crossover rate for each isochore and found that AT-rich isochores experienced less maternal and paternal crossovers than GC-rich isochores, as has been previously observed, though maternal crossovers were sharply diminished in the highest-GC isochores (Fig. 6C, D) [22, 92, 97, 98]. We noticed that these low-crossover, high-GC isochores were often at chromosome ends, where maternal recombination has been shown to be low [99]. To systematically examine recombination relative to each gene along the chromosome, we generated a Manhattan plot of crossover rate for each gene and its flanking regions (Fig. 6B, Additional file 3: Fig. S6A). This highlights the higher recombination of genes located in GC-rich isochores, except for maternal recombination at chromosome ends.

Recombination in paralogous clusters can induce dramatic insertions, deletions, and chromosome rearrangements if paralogues errantly pair with one another [100]. While clusters of outward-looking paralogues are enriched in AT-rich regions, the conserved clusters of paralogues located in GC-rich regions, such as Hox clusters, may also be risky to recombine. To separate the influence of AT/GC content versus clustered-ness, we compared recombination rates across clusters of

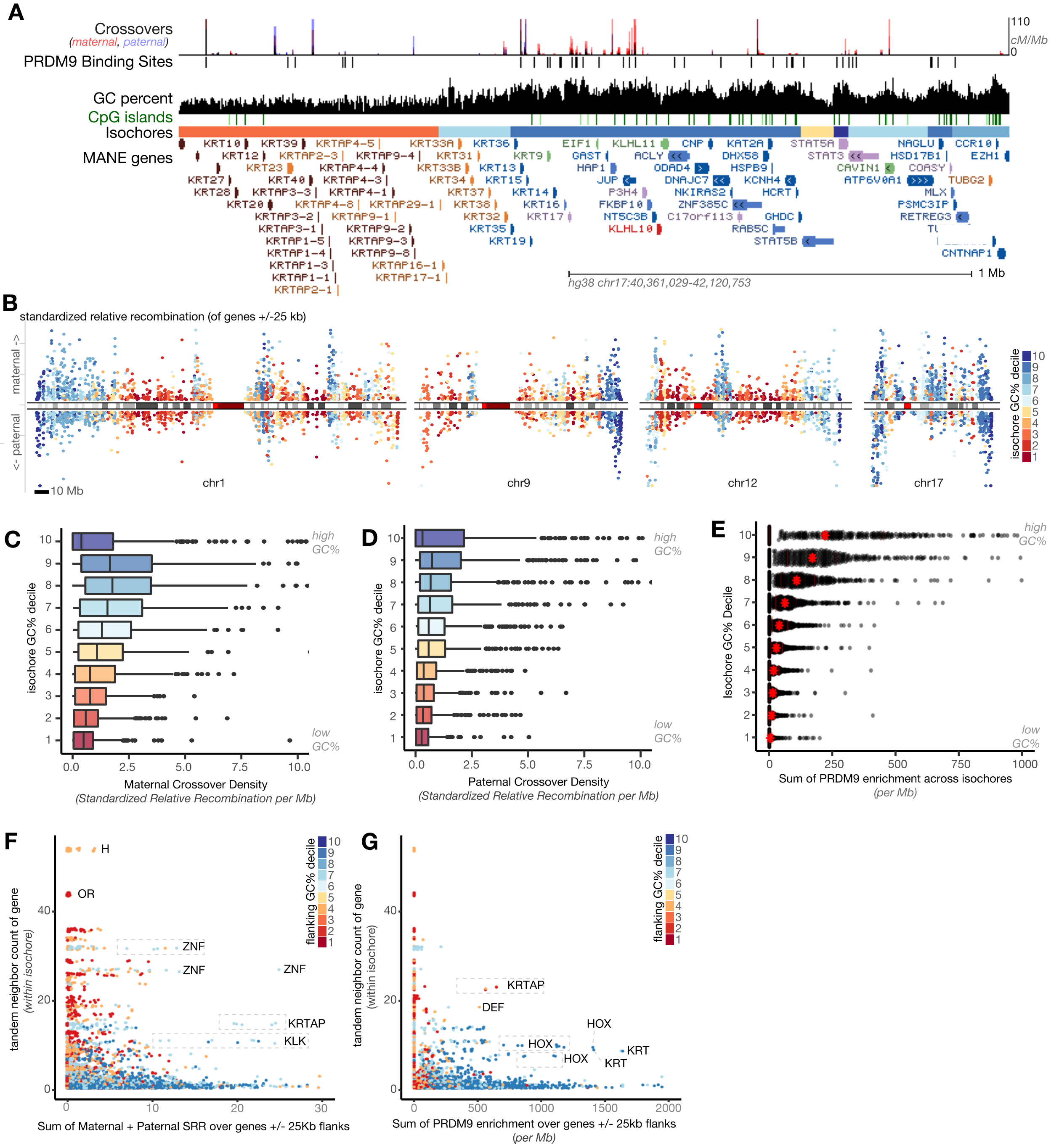

Fig. 6 Crossovers are directed away from AT-rich isochores and genes. **A** UCSC genome browser screenshot of human KRTAP cluster showing recombination rates (maternal meiosis in red, paternal meiosis in blue) and PRDM9 binding calls (black). **B** Manhattan plot of standardized maternal (above 0) and paternal (below 0) recombination rate for each gene with its 25-kb flanking regions. Genes are colored according to the GC content of their home isochore (Red: high AT. Blue: high GC). Scale bars: 10 Mb. Chromosome ideograms reflect centromeres (bright red), gaps (dark red) and Giemsa bands (grays). Recombination rates are low for genes located in AT-rich isochores, except at chromosome ends, which are depleted for maternal recombination. **C** Maternal and **D** paternal crossover rates in isochores binned by GC percent. Crossover calls are from deCODE [92]. **E** PRDM9 peak enrichment across isochores binned by GC%. **F**, **G** Scatterplot of tandem neighbor counts per genes within the same isochore versus **F** the sum of maternal and paternal standardized relative recombination and **G** sum PRDM9 enrichment

different sizes (Fig. 6F). However, most large clusters of genes were AT-rich and had little recombination, making it difficult for us to separate which variable dominates. While recombination is generally directed away from genes, some genes experienced crossovers. We plotted the isochore and *k*-means cluster distribution of genes with the 10% highest internal crossover rate (TSS-TES, Additional file 3: Fig. S6D-E). These genes were in GC-rich isochores and excluded from AT-rich *k*-means cluster 3.3. Together, these analyses suggest that gene blooms located in AT-rich regions of the genome experience low current crossover rates.

In many vertebrates, including humans, crossovers are seeded by binding of PRDM9 to its target site [101–103]. To test whether variation in observed recombination across AT/GC categories is due to differential seeding, we examined PRDM9 binding data from human cells (Fig. 6E, G, Additional file 3: S6F-G) [104]. We observed a striking depletion of PRDM9 binding from AT-biased gene clusters and isochores, in line with previous analyses [97]. We note that many animals have lost PRDM9, and in the absence of PRDM9, recombination is often seeded at CpG islands [105]. As described below, AT-biased gene families also lack CpG islands. Taken together, these results suggest that AT-biased, tandemly arrayed gene clusters experience low rates of recombination and that this likely results from low rates of crossover initiation.

Transcriptional regulation of outward-versus inward-looking genes

Our analyses above lead to the hypothesis that clusters of paralogues that are not subject to purifying selection (i.e., "outward-looking genes") lose GC content as they bloom due to reduced recombination and increased tolerance of point mutations. We further hypothesize that the sequence content and context of these genes has been evolutionarily co-opted to produce extreme tissue specificity and/or variegation of their transcription. While we expect that isochore-level AT/GC content influences genomic organization and histone mark flavors, which would influence transcription indirectly, we also noticed that paralogous clusters of genes lacked annotated CpG islands in their promoters (see CpG Island track, Fig. 2A, B, Additional file 3: Fig. S2D) [66]. CpG dinucleotides are depleted from vertebrate genomes due to the mutability of methylated cytosine; nevertheless, CpGs are relatively enriched in vertebrate promoters, and these "CpG islands" frequently remain unmethylated [106]. Previous analyses suggested that 50–70% of mammalian genes have CpG island promoters [26, 107, 108]. Nevertheless, by considering each gene in its sequence context, we estimate that 90% of protein-coding genes have GC enrichment directly upstream of the TSS (Fig. 3A).

While CpG islands are calculated relative to the GC content of the region, we sought to ask if it is even *possible* to have a CpG island in an AT-rich region [109]. As in Fig. 2, we separated isochores of at least five genes by GC content, gene number, and prefix diversity, and then calculated CpG island scores on a per-gene basis (Fig. 7A, B). Surprisingly, genes located in all types of isochores *could* have CpG islands in their promoters, but genes in large paralogous arrays often did not. Highly conserved genes in arrays, like Hox and Histone arrays, retained their islands regardless of the GC content of the isochores. Gene blooms lacking CpG islands tended to be outward-looking, and almost all large, island-poor arrays were in AT-rich isochores. We also noticed that the shape of islands around the TSS varied across our *k*-means clusters, with some clusters having islands that were symmetrical relative to the TSS, and others having islands that were polarized to the 3′ side of the TSS (Additional file 3: Fig. S7A). CpG island length has also been previously associated with expression heterogeneity [110].

Because CpGs are only protected from methylation when they are located near other CpGs, i.e., in islands, and because methylated CpGs are extremely mutation-prone, CpG loss can be precipitous once it begins [39]. We calculated observed promoter CpG sites compared to the expected rate based on local sequence content and observed a bimodal distribution across genes of all isochores, though the observed/expected CpGs elsewhere in the isochore were predicted by GC content (Fig. 7C, D, Additional file 3: Fig. S7A-B). Thus, though CpG-less promoters are more common in high-AT isochores, islands are compatible with genes in any isochore context [109]; island loss appears specific to paralogous gene blooms (Fig. 7E). Across a range of isochore GC contents, genes near tandem neighbors had CpG-poor promoters compared to singletons (Additional file 3: Fig. S7E). This promoter state also drives *k*-means clustering of genes in 3.1 and 3.3 (Fig. 7D). Many genes in 3.1 and 3.3 have no more CpGs in their promoters than the genome-wide average (~25% of expected CpGs). Further separating these *k*-means clustered genes into those with paralogous neighbors and those without did not stratify promoter architecture, suggesting that the *k*-means categories are already sensitive to the AT/GC sequence features delineating genes in paralogous clusters (Additional file 3: Fig. S7D).

As we and others suggested previously in the mouse, CpG-less promoters are likely to be regulated by non-canonical mechanisms that are independent of TATA-binding protein (TBP) [20, 111]. This could allow the

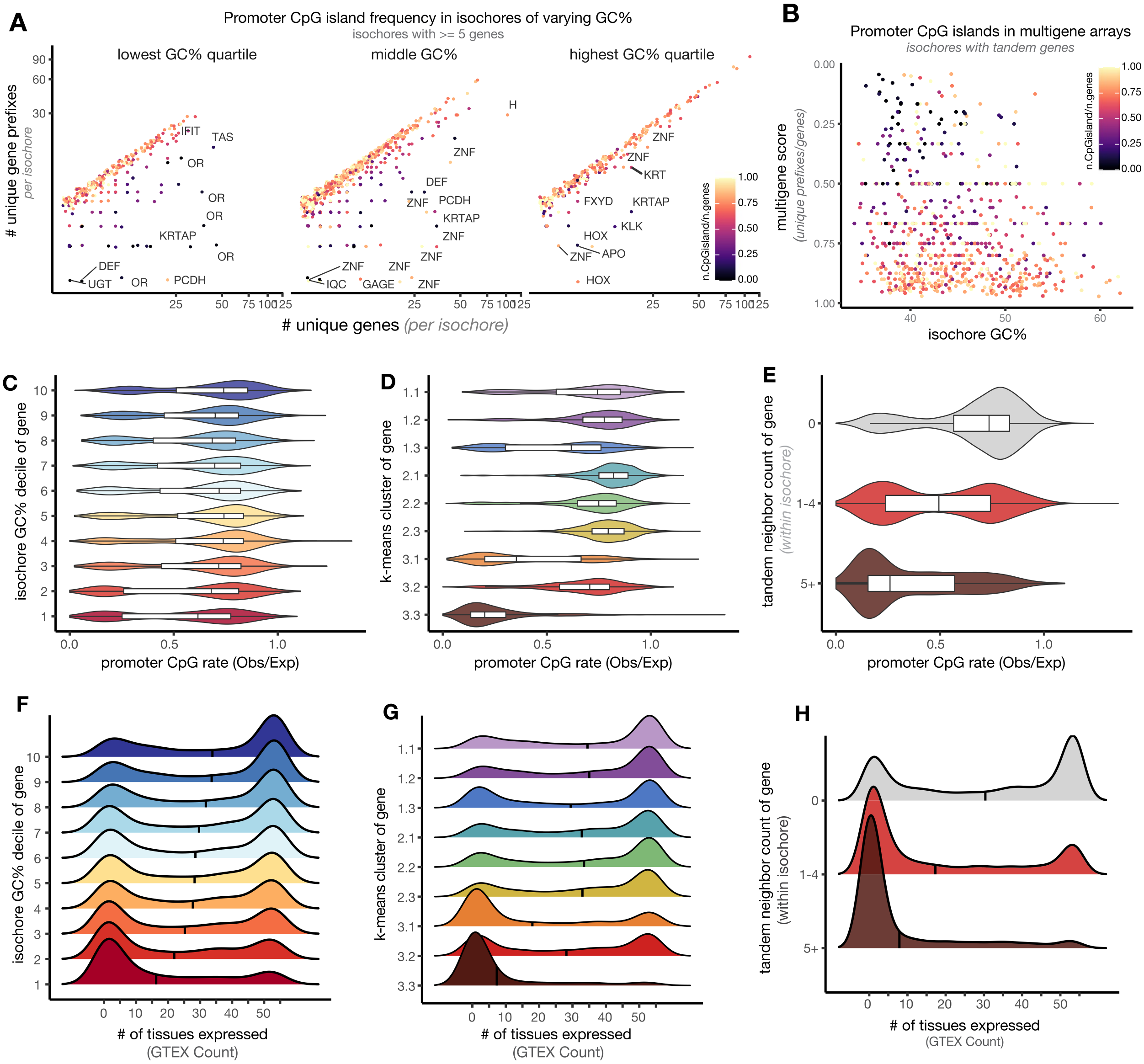

Fig. 7 Lack of CpG islands and restricted expression is a feature of tandemly arrayed AT-rich genes. **A** Promoter CpG island frequency (from [66]) in isochores of varying gene content. Same plots as in Fig. 2E, recolored to highlight promoter characteristics. Most common gene name prefixes in the isochore are labeled. **B** Multigene score versus isochore GC% for isochores containing tandem duplicated genes. Color represents fraction of genes in the isochore with promoter CpG islands. **C**–**E** Distribution of promoter CpG rate for genes binned by **C** isochore GC% decile, **D** *k*-means cluster assignment, or **E** tandem duplication score. We counted CpGs from −750 to +250 bp around the gene TSS and compared to the expected rate for a sequence of that GC content. **F**–**H** Distribution of tissue-level gene expression for genes binned by **F** isochore GC% decile, **G** their *k*-means cluster assignment, or **H** tandem duplication score. GTEx data from 54 human tissues was binarized to "expression" or "no expression" based on RPKM of 5. AT-rich genes and those in paralogous arrays are detected in few or no sampled GTEx tissues, while GC-rich genes are frequently detected in all sampled tissues. Black bars depict medians

unique and rare expression of these genes relative to their CpG-containing brethren: their ground state is to be "off forever". While hemoglobin β and amylase genes, which we now know to be small clusters of paralogues, were shown to lack CpG islands in their promoters in the 1980s, measurements of the relationships between tissue specificity and promoter architecture were performed before the transcription start sites and promoters of the vast majority of tissue-specific genes were mapped [22, 26, 51, 112]. To test how AT/GC distribution in genes relates to patterns of gene expression, we used GTEx data, which measures gene expression in 54 tissue types taken from adult human donor cadavers [113]. We set a threshold (RPKM of 5) to binarize this quantitative data to "expression" or "no expression". In agreement with Bird's hypothesis, this simple metric varies sharply across genes of different *k*-means clusters or with differing promoter or home isochore GC content (Fig. 7F–H) [20, 22, 26, 50, 112]. Genes that are GC-rich are often expressed in many or most tissues tested, while AT-biased genes most often appear to be expressed nowhere or in one tissue. We note the many genes with "no" expression in GTEx are specific to tissues not sampled by GTEx (e.g., olfactory receptor genes in the olfactory epithelium). We infer that genes not detected in any tissue in GTEx data—i.e., the preponderance of AT-rich genes—are highly tissue-, cell type-, time-, or condition-dependent in their expression. As we have shown throughout this study, these exotically transcribed, CpG-less genes are located in evolutionary dynamic clusters of paralogues and tend to be housed in AT-rich isochores.

In the longest-lived cells in the body, post-mitotic neurons, tandemly arrayed gene families have been shown to be clustered with one another in nuclear space and to be uniquely protected from accumulation of CpH (Cp-nonG) methylation [56, 114, 115]. These findings, together with the overwhelming transcriptional repression of these gene families, suggest that they might be sequestered in nuclear space away from the transcriptional machinery. Indeed, we found that across 21 tissues sampled by Hi-C, AT-rich genes, and genes located in AT-rich isochores were likely to be located in transcription-suppressing "B" compartments (Additional file 3: Fig. S7F-G) [116]. Previous analyses have demonstrated that variation in GC content also predicts local chromatin looping and association with the lamina and other nuclear structures [23, 46, 63, 117].

Discussion

Animals make extensive and varied contacts with the external environment, both engaging with foreign molecules and producing and excreting their own substances. Specialization of these input-output functions plays a definitive role in animal lifestyle and often occurs in mammals via amplification and diversification of tandemly arrayed gene families [10, 12, 20, 58, 89]. Extensive gene losses are also common—for example, just as the vomeronasal organ is vestigial in humans, human genes for vomeronasal receptors are no longer functional [19, 118]. Here, we define a common genomic architecture in mammals for genes whose products engage the external world: they are in tandem arrays, are found in AT-biased isochores, and lack CpG islands in their promoters (Fig. 1) [20, 53]. In humans today, we find that genes in AT-rich tandem arrays experience low rates of point mutation and recombination and tend towards tissue-specific expression patterns.

Modeling the emergence of AT/GC sequence bias in outward-looking tandem arrays

The rapid turnover of genes in tandem arrays by birth-and-death evolution can complicate efforts to reconstruct their evolutionary history. Using synteny, we were able to track paralogue number changes for a set of polymorphic arrays in mammals (Fig. 4). For arrays of olfactory receptors, skin proteins, and defensive protease inhibitors, AT content increases with copy number across mammals. This analysis suggests that high AT content need not be an ancestral feature of gene blooms but can emerge during array expansion. Using population genetic data from humans, we test whether elevated rates of allelic diversity in outward-looking genes results from distinct mutational or selective effects. We accept the argument of Lynch that AT/GC content itself is unlikely to be directly operated on by selection, and instead focus on neutral and selective events affecting the genes contained in sequence-biased isochores [34].

We find that genes in AT-biased tandem arrays experience low ongoing rates of point mutation (Fig. 5) and low rates of recombination (Fig. 6). We suggest that excessive allelic diversity in these regions could have arisen due to weakened selection on historical point mutations. Because the most common point mutations replace C or G bases with T or A bases, tolerance of point mutations in expanding tandem arrays could shift GC content down; low rates of point mutation in the present would be due to the scarcity of mutable CpG dinucleotides remaining in these clusters [45]. Tandem arrays also appear depleted for recombination events that can shift GC content backup (Fig. 6). Together, tolerance of historical point mutation and strengthened intolerance of recombination as gene families expand would tend to bias expanding arrays towards higher AT content (Fig. 1).

While deamination of cytosine in the CpG context dominates mutation rates, many other neutral processes contribute to the overall mutation spectrum, and these too operate differently in copy-number-variable tandem arrays. For example, AT-biased regions of the genome are late-replicating, which is associated with higher rates of germline mutations, and genes in tandem arrays tend not to be transcribed in the testis, which would exempt them from transcription-coupled repair [49, 119]. Excess allelic variation could also result from paralogous gene conversion. While paralogous gene conversion can be measured for paralogous pairs, in extended arrays this phenomenon would be difficult to distinguish from copy number variation without long-read data spanning the region [120]. Similarly, genes in tandem arrays do not have one-to-one homology relationships across species. Trivially, this means that their interspecific divergence is high, but they would be left out of gene-wise measures of divergence that rely on clear homology assignments.

Other mutagenic and selective forces may also contribute to the evolution of AT-sequence bias. For example, the repetitive nature of these regions of the genome predisposes them to problems, such as replication fork slippage, that require non-homologous end joining (NHEJ) to resolve [121, 122]. RCC4-DNA ligase IV is primary ligase complex for NHEJ, and it preferentially ligates poly-dT single-stranded DNA and long dT overhangs [123]. Polymerase mu, which is also involved in NHEJ, shares a similar bias towards both dT and dC, though preferentially towards dT [124]. These biases could encourage these regions of the genome to adopting AT-rich sequences with each duplication, and thus each addition of repetitive material.

Implications for gene regulation

Genes located in AT-biased tandem arrays are typically silenced almost everywhere in the body and expressed at extremely high levels at a specific place and time. Where and when these genes are expressed, they perform the definitive work of the cell type. Many of these outward-looking gene families exhibit some kind of exclusive expression, from the fetal to adult switch in hemoglobin β expression (i.e., exclusion over time) to the one-receptor-per-neuron pattern of olfactory receptor expression (i.e., exclusion over space). We expect that genes located in high-AT isochores that lack CpG islands in their promoters are durably silenced by particular mechanisms when they are not expressed (i.e., almost everywhere), and that their expression will be activated by non-canonical mechanisms in the single condition where each is expressed. As we and others suggested previously for mouse ORs, CpG-less promoters are likely to be regulated by non-canonical mechanisms that are independent of TATA-binding protein (TBP) and the recently discovered basal CpG-island-binding factors BANP and BEND3 [20, 111, 125, 126]. As Bird hypothesized presciently in 1986, “When activated, they appear to be bound to a complex of tissue specific factors which presumably accomplish what…islands can achieve using ubiquitous, non-tissue specific factors. Given that there are usually rather few CpGs near tissue specific genes…one would not expect to find CpG or methylation built into the activation mechanism of genes of this type” [112]. One could argue that these genes do not have promoters at all and rely completely on locus control regions to concentrate and deliver transcription factors to the TSS [54, 127, 128]. This atypical architecture appears to predispose these genes to restricted expression relative to their CpG-containing brethren: their ground state is to be “off forever”.

Many or most molecular genetic events are sensitive to variation in AT/GC distribution: AT content predicts compartmentalization of the genome in 3D space, replication timing, and patterns of histone marks [48, 129–133]. Typically, AT-biased sequence is packaged as heterochromatin and silenced. Work on olfactory receptors, clustered protocadherins, and secreted liver proteins suggest that these gene families are expressed from the context of constitutive heterochromatin, which appears to be present prior to expression and to be retained on family members that are not expressed [134–138]. CpG islands also function as molecular beacons: they mark transcription start sites, serve as recombination hotspots in the absence of PRDM9, and act as replication origins in meiosis [105, 129, 139]. The accrual of high AT content in gene arrays and the lack of CpG islands is therefore likely to exert a strong effect on the molecular regulation of these genes. In ectodermal development, single-copy genes accrue CpH methylation, perhaps passively, while AT-rich gene arrays remain devoid of this modification [114]. This suggests that AT-rich gene arrays are locked away from the ambient molecular stew of the nucleus, perhaps over very long developmental time periods. Indeed, tandem arrays clump together in the nucleus in post-mitotic neurons [56, 115].

In addition to the extreme time and/or tissue specificity of most outward-looking gene families, a fraction of these families exhibit stochastic expression such that each cell expresses just one or a sparse subset of family members. Chemosensors, B- and T- cell receptors, NK cell receptors, and clustered protocadherins all exhibit this restricted expression [136, 140, 141]. As we argued recently, sparse cell-wise expression patterns compartmentalize the effects of mutations [136]. These mechanisms are also likely to result in insensitivity to copy number variation, as each cell chooses its own dose of family members for expression. Feedback mechanisms that ensure cells can “choose again” if they originally pick a pseudogene further buffer potential deleterious effects of mutations [142, 143]. Combined with the sheer numbers of family members that a particular gene function is distributed across, these expression mechanisms could predispose these genes to selective drift.

While the α and γ families of clustered protocadherins are embedded in AT-rich sequence and all three families exhibit variegated expression, these genes depart from the pattern of CpG island loss exhibited by outward-looking tandem arrays: Each clustered protocadherin variable exon promoter is flanked by prominent CpG islands, and cytosine methylation plays a critical role in their transcriptional regulation [144, 145]. We noticed that singleton protocadherins, as well as other small clusters of neurodevelopmental and synaptic genes (i.e., *SLITRK*, *GABR*, and *CDH* families), shared this pattern. A subset of these smaller clusters show a specific dependence on *POGZ* for transcriptional activation [146]. We plan to address the evolutionary history and regulatory consequences for this flavor of gene in a future publication.

Role for recombination in diversification versus maintenance of gene arrays

Local sequence diversity and recombination rates are usually positively correlated (reviewed in [147]). In contrast, we find that AT-biased gene families exhibit high diversity and divergence despite low recombination. Comparison of the human and chimpanzee genomes found two distinct types of high-divergence regions: Genome-wide, chromosome ends are GC-rich and have high divergence and recombination; in chromosome interiors, however, recombination is more modest overall, and divergence is higher in G-bands, which are AT-rich [27]. As Holmquist argued previously based on much more limited data, we show here that these G-bands contain very different types of genes than do Q-bands [22, 57].

Overall GC content is theorized to be a record of past gene conversion, with high-GC regions having experienced high historical recombination [38]. If this is the case, then AT bias in tandem arrays could reflect historical depletion of recombination in these regions, while also dampening recombination rates in the present (Fig. 6). Modeling suggests that once variation in AT/GC content starts to emerge, it can be self-reinforcing via positive feedback [39]. There remains conflict between the mode by which these gene arrays are thought to have bloomed—i.e., via gene duplication through ectopic exchange during recombination—and their current depletion for recombination events. Other modes of duplication, including replication slippage and transposition, may also be at work in expanding these arrays, and duplication mechanisms could themselves influence GC content [148].

Ectopic exchange in repetitive gene regions can have benign or catastrophic consequences. Induction of copy number variation within a gene array may be of small phenotypic consequence, as the jobs these genes perform are by nature distributed across many family members. In contrast, ectopic exchange that deletes a cluster or induces recombination between clusters can result in catastrophic chromosome rearrangements [100]. Retrotransposition or ectopic exchange mediated by repetitive elements can seed new gene clusters elsewhere in the genome [149, 150], and gene families with more than one cluster genome-wide are likely to be particularly dangerous for genome stability. Indeed, mammalian chromosome evolution appears to have been shaped by ectopic exchange between OR clusters, to the extent that ORs are often positioned near chromosome ends [151–156]; recurrent translocations between OR clusters continue to occur [151–153, 157–160]. Finally, even if structural variation in outward-looking tandem arrays is benign within an individual, it can lead to hybrid incompatibility and can initiate or reinforce reproductive isolation that leads to speciation [161–163]. Recent modeling work has sought to characterize the tradeoffs between the structural fragility of gene blooms and the potential positive effects of allelic diversification [164].

Given the genomic danger of these tandem arrays, why have gene family members remained in cis with one another? An extreme example is the “milk and teeth” locus on human chromosome 4. The casein genes in this locus evolved via tandem duplication of enamel genes at the root of the mammalian tree; the enamel genes themselves evolved from *follicular dendritic cell secreted protein* in bony fish [12, 165]. Astonishingly, these genes have remained syntenic. Why would this be the case, given that they are expressed in three separate body systems and that such tandem arrays are genomically dangerous? We propose that as in the case of maintenance of Hox gene synteny, the regulatory elements of these genes remain tangled with one another, such that relocation of array members elsewhere in the genome would divorce them from cis-regulatory elements that they depend on for expression [166–168]. Recent research on enhancer evolution in animals suggests that enhancer tangling can result in the preservation of synteny over ~700 million years [169]. In other cases, as in B Cell Receptor, hemoglobin, clustered protocadherin, interferon, and chemosensor arrays, family members share and compete for the same regulatory elements [128, 170–173]. This mutual dependence would again increase the phenotypic consequences of recombination events that break synteny by separating genes in large families from locus control regions they depend on for expression.

Array incompatibility between individuals of a species and the necessity of remaining co-located with regulatory elements that may be tangled with or shared by other gene family members could cause tandem arrays to behave like supergenes—multigene regions inherited as an allelic unit. We suspect that depletion of CpG islands and PRDM9 sites from tandemly arrayed genes protects the genome from the danger of errantly recombining these duplicative regions. Nevertheless, recombination and gene duplication or deletion still sometimes occur in these regions—their crossover rate even today is nonzero—and the marginal fitness effects of resulting copy number variants could allow products of these meiosis to be preserved in the population. As gene arrays get larger, point mutation tolerance could shift their GC content downward, putting the brakes on recombination as they become ever more unwieldy. Overall, recombination in these regions would be dampened, while differential tolerance of local duplications versus gross rearrangements could allow an increase in local allelic diversity.

Implications for chromosome organization

Repetitive elements have shaped chromosomal evolution since the dawn of eukaryotes. The linear genome is proposed to have arisen from erroneous meiotic recombination between Group II introns which invaded the circular genome to create the t-loop precursors to stable telomeres [174]. Similarly, dispersion and expansion of ORs and other large tandem gene arrays have shaped mammalian chromosome evolution. Tandem arrays of ORs represent ancestral breakpoints of chromosomal synteny between mice, rats, and humans [156, 175]. A large OR cluster is found at the end of the q-arm of chr1 in humans but not in mice. In addition to ORs, large gene families including zinc finger (*ZNF*) and immunoglobulin heavy chain (*IGH*) genes are observed at chromosome ends across eukaryotes [176]. In the modern human population, unequal crossovers between OR clusters are a source of recurrent and pathological rearrangement hotspots [158].

While we also observe these AT-rich isochores *near* chromosome ends, terminal isochores are often some of the most GC-rich in the genome and can contain heterogeneous single-copy genes [177]. Indeed, the largest OR cluster at the end of the q-arm of chromosome 1 in humans is followed by a higher GC% isochore containing *ZNF* genes. This strong end-GC% accumulation appears to arise paternally: genes in high GC% isochores at chromosome ends are enriched for paternal crossovers and relatively depleted of maternal crossovers. Overall, paternal crossovers are biased towards chromosome ends [99, 178]. Chromatin organization of pachytene spermatocytes is implicated in this phenomenon. Spermatocytes have shorter synaptonemal complexes compared to oocytes, and, further, subtelomeric regions do not require PRDM9 for crossovers; however, how these factors contribute to paternal crossover end-bias remains incompletely understood [179, 180]. Potentially, recombination-based alternative lengthening of telomeres (ALT) in spermatocytes biases hotspots towards chromosome ends [181].

Over evolutionary time, as ectopic recombination places high-AT tandem arrays at chromosome ends, high paternal rates of gBGC at the ends of chromosomes would generate new isochores of increasing GC% comprising newly evolving genes [182, 183].

Is mutation biased or random with respect to gene function?

Recent mutation accumulation studies have suggested that de novo mutations could occur with different frequencies in different kinds of genes or in genic versus non-genic locations [184]. We and others also argued previously that segregation of mutation-tolerant versus mutation-intolerant genes into AT- versus GC-biased regions of the genome could allow differential mutation rates on different classes of genes [20, 80, 81]. However, our analyses of synonymous versus non-synonymous variants and of de novo mutation rate in AT- versus GC-biased genes suggest the opposite: that AT-biased genes experience fewer mutations in living humans than do GC-biased genes. This is in line with neutral, sequence-based expectations [49]. While CpG prevalence dominates the mutation spectrum, other mutational and repair processes that differentially affect AT- versus GC-biased regions of the genome almost certainly contribute to the overall pattern of diversity in inward- versus outward-looking genes. In particular, AT-rich regions of the genome are late-replicating, and tandemly arrayed genes are excluded from transcription-coupled repair during spermatogenesis. Each of these patterns could increase the mutation rate in outward-looking genes. Late replication can worsen the loss of cytosines, while lack of transcription-coupled repair would be expected to increase the rate of A>G (T>C) mutations [49, 119].

While mutagenic or repair processes specific to gene blooms could contribute to their overall mutation spectrum, we expect that for most, allelic diversity and differential AT/GC content in inward- versus outward-looking genes result from differential selection trajectory over evolutionary time. One attractive model is that as a gene cluster expands in size and the function of that gene family is partitioned over more and more members, purifying selection becomes weaker on individual family members (Fig. 1). Because the most common point mutations are C->T and especially CpG->TpG, evolutionary tolerance of point mutations would shift GC content down. On the other hand, purifying selection and higher recombination rate would both preserve the GC content of singleton genes.

Matching sequence content for genes and their isochore environments

Throughout this study, we show that the local AT/GC content of genes correlates closely with that of their isochore environment (e.g., Fig. 3B, Additional file 3: Fig. S3B-G). We found this pattern perplexing. For example, while higher purifying selection on single-copy genes can help to explain their higher GC content, we would expect most variants in the non-coding portions of an isochore to be phenotypically neutral. Because gene conversion acts over longer chromosomal distances, differential recombination could affect the sequence content of genes and their environment together. This pattern could also result from background/linked selection, which is known to vary in strength across the genome [185, 186]. Because function-changing variants are physically linked to local variants that do not alter phenotype, differential purifying selection on mutations in single-copy versus multicopy genes would affect the strength of background selection in the neighborhood and inheritance of linked neutral variants. Background selection will operate over shorter distances in regions with higher recombination (i.e., GC-rich isochores), but because multiple inward-looking, single-copy genes are co-located in the same isochore, there may be little nearby that is exempt from linked selection. In this way, purifying selection that results in maintenance of high GC content in single-copy genes would also preserve the high GC content of their isochore neighborhood. Finally, variation in repetitive element distribution could also contribute to the AT/GC content of isochores housing outward-looking versus inward-looking genes. Indeed, increased LINE density in chemosensory gene families has been proposed to contribute to their regulation [187].

The matching sequence content of genes and their isochore neighborhoods would seem to facilitate coherent patterns of histone marks across entire isochore units and their organization in nuclear space. On the other hand, genes with AT- versus GC- rich coding sequences use distinct codons, which could affect their translation rates given variation in the prevalence of various tRNAs. Further analysis will be required to assess whether amino acid distribution varies for proteins coded by AT-rich versus GC-rich sequences.

Finally, while we can find genes with GC-rich promoters in all kinds of isochores (Additional file 3: Fig. S3F), we never find genes with AT-rich promoters in GC-rich isochores. While we favor the model shown in Fig. 1, where isochore-level AT content rises during gene blooms due to the parallel actions of protein-coding drift and suppressed recombination, we cannot exclude (and are indeed fascinated by) the possibility that the distinct regulatory characteristics that allow LCR-mediated expression could be the keystone factor tying outward-looking genes to AT-rich isochores. For example, reduced expression of duplicated genes or lack of CpG islands have been associated with more durable evolutionary retention of duplicates, though these studies focused on paralogous pairs and patterns may be different in blooms [188, 189]. Overall, in the absence of other evidence, we think that LCR-driven transcription is a kludge that allows expression of paralogous clusters that failed to maintain their CpGs, rather than CpG-less promoters emerging via positive selection to allow regulation by LCRs.

Is this genomic architecture specific to mammals?

While isochore structure is not unique to mammals, it is not a universal feature across animal clades, and the AT/GC variation observed in mammals is extreme [34]. We are curious whether stem mammals evolved molecular mechanisms that facilitated the evolution of gene arrays. These could include both systems that maintain these arrays as constitutive heterochromatin when they are not being expressed and unique transcriptional mechanisms that activate them, often in a stochastic or highly restricted manner, in their target tissues. One candidate factor that mediates long-range enhancer-promoter interactions in multiple arrayed families is Ldb1 [54, 190]. Social insects have also massively expanded their olfactory receptor gene repertoire in cis; in the ant, this is accompanied by increased AT content [191]. Have convergent mechanisms for stochastic expression facilitated olfactory receptor repertoire expansion in insects?

Other clades may have evolved distinct mechanisms to organize repetitive genes or gene pieces: In *Diptera*, repetitive arrays are often organized as alternative splicing hubs [192–195]. Reptiles and birds exhibit “microchromosomes” which have distinct GC content from the rest of the genome and can house arrays of rapidly evolving, outward-looking genes such as venoms [196]. Trypanosome arrays of surface VSGs are located in subtelomeric regions [197]. For mammals, the “isochore solution” balances diversity in gene arrays with genomic integrity.

Conclusions

To date, models of mammalian genome evolution have not included both the sharply distinct functions and transcriptional patterns of genes in AT- versus GC-biased regions. Here, we describe a common genomic architecture in mammals for genes whose products engage with the external world, which we call outward-looking genes: these genes tend to occur in tandem arrays, occupy AT-biased isochores, and lack CpG islands in their promoters (Fig. 1) [20, 53]. In addition, we find that in humans today, genes in AT-rich tandem arrays experience low rates of point mutation and recombination and tend towards tissue-specific expression patterns. We hypothesize that as a gene family expands in cis, selection on amino acid sequences of individual family members weakens, while selection against recombination grows. Together, these forces result in loss of GC bases over evolutionary time.

Once AT-rich, a tandem array is more likely to be heterochromatinized, allowing it to acquire highly tissue-specific expression patterns, a feature which is enhanced when combined with the loss of CpG islands. Variegated expression reduces the phenotypic consequences of change in copy number, allowing further rounds of copy number change, which provide an organism with the flexibility to adapt to an ever-changing environment.

Methods

Describing isochores

To call isochores, we implemented a genomic segmentation algorithm called GC-Profile [64] using halting parameter (number of segmentation iterations) of t_0 275 and minimum segment length of 3000 bp. Gaps less than 1% of the input sequence were filtered out, generating 4328 distinct isochores in hg38 (Additional file 1). Isochores were ranked by average GC%, with rank 1 having the highest and 4328 having the lowest. We also performed this analysis in hg19 (Additional file 2).

Statistical analyses

We performed Kruskal-Wallis non-parametric ANOVA for each group of comparisons (Additional file 4). We then used Dunn's pairwise analysis to compare individual groups with one another (Additional file 4). We use phylogenetic generalized least squares (PGLS) regression as implemented in caper [198] to test for correlations between genomic characteristics across vertebrate species while controlling for the evolutionary non-independence of the multi-species dataset. We used timetree.org [199] to produce our species phylogeny for these analyses, where the tree was imported and pruned to the subset of species for a given comparison using the ape package [200]. In PGLS, lambda was optimized by 0 and 1 by maximum likelihood.

GC content calculations

Genes from the Matched Annotation dataset from the NCBI and EMBL-EBI (MANE) Select dataset [67] were downloaded from the UCSC Genome Browser. Isochores were matched to genes using the coordinates of the transcription start site. GC content across gene features, including promoters (−750 to +250 bp flanking TSS), flanking regions (+/−25 kb), coding exons, exons and UTRs, and introns were separately calculated from FASTA sequences using bedTools [201].

To generate 9 k-means clusters, we used gc5BaseBw from the UCSC Genome Browser [65] to calculate GC% scores across MANE genes with +/−1 kb flanks. We generated 3-kmeans clusters of genes, which were further clustered into 3-kmeans clusters each using deepTools plotHeatmap [202]. Cluster assignment and quantification of other parameters for each gene are reported in Additional file 5.

Characterizing types of genes

To characterize the types of genes residing in isochores of varying GC, we used 2 categories of descriptors: GO terms and gene prefixes. To identify GO terms associated with genes in each isochore GC decile, we used the R package, clusterProfiler (version 4.2.2) [203]. This helped us streamline identification of key terms that appeared in each decile. With this list, we identified GO terms that were most significantly enriched in each decile with a depth of at least 30 genes. Using AmiGO [204], an online database of GO identifiers, we pulled the list of genes associated with our selected group of significant GO terms and plotted GC content across each term. The terms we chose are listed in the table below.

Shortened term (from Fig. 2**)**	**Full GO term description**	**GO ID**
wnt signaling	wnt signaling pathway	GO:0016055
kinase activity	kinase activity	GO:0016301
transcription	transcription, DNA-templated	GO:0006351
defense	immune response	GO:0006955
xenobiosis	xenobiotic metabolic process	GO:0006805
keratinization	keratinization	GO:0031424
chemosensation	detection of a chemical stimulus	GO:0009593

We wanted an alternative to GO analysis for assessing diversity across isochores and *k*-means gene clusters. Since the prefixes of well-annotated genes (like the ones from the MANE dataset) are shared across genes within the same gene family, we used this as a means of assessing diversity with more specificity than one would achieve through GO analysis. The process of assigning gene prefixes is as follows:

(1) Convert old names into new nomenclature.

- Go to the HUGO Gene Nomenclature Committee's (HGNC) [205] website and the list of gene symbols from the MANE set into their "Multi-symbol checker". This will ensure we have the most up-to-date names for each of our genes (i.e., some which may have been labeled as "FAM" may have a new symbol to go with the rest of the gene family).

- Match names in the MANE set to names labeled “Approved symbols” by HGNC, and replace those symbols with the HUGO names.

(2) Replace numbers with “_”. We cannot remove all numbers because there are several genes that have more letters after numbers that are not important for our purposes (i.e., CSN2A will become CSN_A).
(3) Remove anything after the first instance of “_” (i.e., CSN_A will become CSN). The goal of this step is to keep the first part of the prefix, removing letters and numbers that indicate subfamilies.
(4) While it is not common, some genes require us to know those numbers to know what they do (most commonly, enzymes involved in modifying carbohydrates). Largely, these genes start with a single letter, followed by numbers, then more letters. Thus, to fix these genes, we pull out the genes that have 1 letter after steps 4 and 5.
(5) Look through each of those genes that start with only one letter, then decide how best to group them.
(6) View gene prefixes in alphabetical order and search for prefixes that are likely to be families, then rename (i.e., KCNT and KCNQ are both potassium channels, so we grouped these together).

Once we had a list of gene prefixes, we calculated a Shannon’s H diversity metric for each isochore based on the prefix probabilities in each isochore (proportions + log2(1/proportions) = diversity metric). Larger values are indicative of more diversity. Similarly, we calculated a Shannon’s H diversity metric for each *k*-means cluster.

De novo mutations

De novo mutations (DNMs) were compiled by [85] from seven family-based whole genome sequencing (WGS) datasets, encompassing a total of 679,547 single-nucleotide variants (SNVs), which comprise data from both neurotypical and neurodivergent individuals. We remapped the dataset to hg38 using LiftOver in UCSC Genome Browser. To calculate genomic DNM density, we counted the number of DNMs occurring within the coordinates listed in the GC calculation section above. To calculate DNM density, we pooled genic DNMs within each isochore and divided by the sum of the region of interest’s size, i.e., we identified all the genes in an isochore, summed the DNMs between their transcription start and termination sites, then divided by the summed length of those genic regions.

For analysis of human genetic variation (DNMs, allelic variation, and recombination), we checked whether any of the tandem arrays we analyzed were too duplicative to allow mapping of short-read data or blacklisted for other reasons. We found that these genes were sufficiently different that these regions are rarely blacklisted. We compared our regions of interest to “problematic regions” compiled on the UCSC Genome Browser. These encompass “ENCODE Blacklist” regions, which are problematic for short-read sequencing; “GRC Exclusions,” i.e., regions known to be incorrect in the hg38 reference but not yet removed; and “UCSC Unusual Regions,” which are regions that cause frequent confusion for other reasons, such as due to there being segregating alternative haplotypes. Our outward-looking tandem arrays were almost never flagged. We noted two cases that were flagged: A portion of the UGT2 array, centered around *UGT2B17*, which is flagged because there are distinct segregating haplotypes [13] and in the clustered protocadherin locus, which is indeed too duplicative to obtain good mapping of short reads. These two cases will not affect our genome-wide analyses. The BCR and TCR loci also meet UCSC criteria to be confusing for short-read sequencing, but we did not include these loci in our quantitative gene-wise analyses as the repetitive segments are gene pieces rather than independent transcribed units.

Allelic variants

We used the gnomAD v2.1.1 dataset of single-nucleotide allelic variants [82]. The authors defined rare single-nucleotide variants (< 0.1% allele frequency) from 125,748 exomes and 15,708 whole genomes and predicted whether variants within coding regions are likely to be functionally synonymous, missense, or loss-of-function. Here, we used observed synonymous, missense, and loss-of-function mutation rates. We ported variant calls to MANE genes in hg38 using the gene symbol and Ensembl transcript IDs. In Fig. 5, we also use the calculated pLI score from gnomAD, which describes the likelihood that a gene is loss-of-function intolerant in humans.

Recombination

Crossover data for hg38 was acquired from deCODE where the authors used whole genome sequence (WGS) of trios and were able to refine crossover boundaries for 247,942 crossovers in 9423 paternal meioses and 514,039 crossovers in 11,750 maternal meioses [92]. Of note, the data we used here is restricted to autosomes. To calculate crossover density, we assigned crossovers to a region of interest based on the median of the crossover coordinates. We normalized counts within a region by dividing by the genomic average for that sex. PRDM9 binding data from HEK293T cells transfected with the PRDM9 reference allele was acquired from [104]. We selected the top 10% of PRDM9 peaks based on enrichment scores to account for weak PRDM9 binding sites associated with overexpression in the system, as noted by the authors. Like with crossovers, we calculated PRDM9 binding site density across genes as the summed enrichment scores across genic regions within an isochore, mapping by the midpoint of the binding coordinates.

Gene regulatory information

To determine tissue specificity of gene expression, RNA-sequencing data was sourced from Genotype-Tissue Expression (GTEx) project (V8, released in August 2019), containing 17,382 samples collected from 54 tissues from 948 donors [113]. For each gene in the MANE set, we counted the number of tissues in which expression was at least 5 transcripts per million (TPM).

To measure A/B compartment occupancy of genes across tissues, AB compartments were sourced from published Hi-C data from 21 tissues and cell types [116]. MANE genes were lifted over into hg19 to match A/B compartment domain calls in hg19. Isochores called in hg19 were assigned to a compartment by matching the isochore's midpoint to the midpoint of the closest compartment. Genes were assigned to a compartment by matching the transcription start site to the midpoint of the nearest compartment, as most genes did not fall into a single compartment (~90%). Further, we counted the occurrences of compartments A and B for each isochore and gene. These counts were binned into always A (21 counts of A), mostly A (14–20 counts of A or 0–6 counts of B), equally A or B (7–13 counts of A or B), mostly B (0–6 counts of A or 14–20 counts of B), and always B (21 counts of B).

To identify genes with CpG islands in promoter regions, we downloaded the CpG Island track (unmasked) from the UCSC Genome Browser [66]. The ratio of observed vs. expected CpG dinucleotides was converted to a bigwig coverage file and plotted across gene TSS's (+/−2.5 kb) in 9 k-means clusters using deepTools plotHeatmap. The average score of CpG islands within −750 bp and +250 bp of a gene TSS were calculated using bedTools.

Gene Bloom Evolution with Tandem ClipR

To access and analyze syntenic tandem gene blooms across species, we created a pipeline we call TandemClipR that defines tandem arrays based on orthologous "bookend" genes. Bookends define syntenic array bounding positions and are not members of the bloomed family. First, we used the biomaRt R package [206] to define the human genome as the reference mart, and then created a table of bookending ortholog chromosome names, starts, and stops using the "getLDS" function and the dec2021 Ensembl archive as a source. We tabulated these ortholog positions in all 193 vertebrate species in the archive (See Additional files 7 and 8). If both orthologs were on the same scaffold, indicating a contiguous assembly of the tandem array, we used the retrieved coordinates to subset the gff3 annotation file to the focal region using the subsetByOverlaps() function in the GenomicRanges package [207]. Filters were applied when imported annotations to retain only "gene", "mRNA", "CDS", "exon", "five_prime_UTR", and "three_prime_UTR" annotations.

The resulting gff3 files were used to analyze feature GC content across each tandem array based on sequences in the Ensembl "toplevel" assemblies. Namely, we used command-line tools to subset the gff3 annotations into focal features (genes, exons, five_prime_UTR), and used while awk and bedtools merge and subtract [201] were used to produced genic, intergenic, exonic, intronic, and whole cluster bedfiles defining these regions. Additionally, bedfiles defining the promoter region of each gene were produced by defining a window 750 bp upstream and 250 bp downstream of the transcription start site. We fed these bedfiles to bedtools nuc to calculate GC content for each feature, which were counted and aggregated into average GC% for final analyses. We spot-checked gene counts in NCBI's Genome Data Viewer and noticed that there were a few cases in which counts varied between recent assemblies. For example, the lion HBB-OR cluster is much shorter in the assembly we used compared to the most recent assembly. However, our manual counts of the HBB-OR array in a subset of species produced similar trends as the automated annotation. To further verify that one of these assembly/annotation errors did not sway the overall trend, we ran each PGLS analysis iteratively, removing one species at a time, and found that PGLS *p*-values of the whole were largely reflected in each iteration.

Because a recent analysis has suggested that the highest-GC regions can be missing from lower-quality genomes [208], we repeated our analyses on a subset of genomes, including only recent assemblies templated on long-read sequencing (~1/4 of our full set). While this reduced our statistical power for PGLS, as some clades of animals were missing from the higher-quality assemblies, the trends were identical (data not shown).

Abbreviations

Mb	Megabase
Kb	Kilobase
bp	Base pair
gBGC	GC-biased gene conversion
TSS	Transcription start site
TES	Transcription end site
UTR	Untranslated region

GO	Gene ontology
CDS	Coding sequence
OR	Olfactory receptor
EDC	Epidermal differentiation complex
PGLS	Phylogenetic least squares
CpG	Reading 5′ to 3′, a cytosine positioned next to a guanine
$C^{me}G$	Reading 5′ to 3′, a methylated cytosine positioned next to a guanine
CpH	Reading 5′ to 3′, a cytosine positioned any base other than guanine
pLI	Loss-of-function intolerance
DNM	De novo mutation
PRDM9	PR domain zinc finger protein 9, a protein responsible for seeding recombination
cM	Centimorgan
RPKM	Reads per kilobase of transcript, per million mapped reads
NHEJ	Non-homologous end joining
TCR	Transcription-coupled repair
LCR	Locus control region

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-023-01673-4.

Additional file 1. Isochore calls in the hg38 human genome assembly.

Additional file 2. Isochore calls in the hg19 human genome assembly.

Additional file 3: Supplemental figures related to main figures 2-7.

Additional file 4. Results of all statistical analyses presented here.

Additional file 5. Gene-wise data, including isochore assignment, k-means cluster, local sequence characteristics, variant rates.

Additional file 6. Description of data in Additional file 4.

Additional file 7. GC contents of selected gene families and their annotated features (ie: CDS, introns, intergenic regions).

Additional file 8. List of genomes searched with Tandem ClipR.

Acknowledgements

Thanks to Rachel Duffié, Sebastian Zöllner, David Lyons, Steven Clark, and members of the Clowney lab for discussion and comments on the manuscript. This work was supported by the Rita Allen Foundation Milton Cassel Scholarship and the Alfred P Sloan Research Scholarship in Neuroscience to EJC. EJC is a McKnight Scholar and a Pew Biomedical Scholar. MVB was supported by the NIH Cellular and Molecular Biology Training Grant (T32-GM007315) and an F31 from NINDS (F31-NS127484). MAC was supported by the NIH Early Stage Training in the Neurosciences Training Grant (T32-NS076401) and by the Hearing, Balance, and Chemical Senses Training Program (T32-DC000011).

Authors' contributions

All authors read and approved the final manuscript. Conceptualization, MVB, MAC, EJC; Methodology, MVB, MAC, MLH, EJC; Investigation, MVB, MAC, MLH, EJC; Formal Analysis, MVB, MAC, MLH; Visualization, MVB; Data curation, MAC; Writing—Original Draft, MVB, MAC, EJC; Writing—Review and Editing, MVB, MAC, MLH, EJC; Funding Acquisition, EJC; Supervision, EJC.

Availability of data and materials

All data and analyses described for this study are either included as additional files listed below or are available on GitHub. Scripts for tandemClipR can be found at https://github.com/Matthew-Holding/TandemClipR [209]. UCSC Genome Browser tracks used for visualization throughout this paper can be viewed at http://genome.ucsc.edu/s/mbrovkin/hg38 [210].

Declarations

Ethics approval and consent to participate

NA

Competing interests

The authors declare that they have no competing interests.

Received: 25 April 2023 Accepted: 1 August 2023
Published online: 23 August 2023

References

1. Charkoftaki G, Wang Y, McAndrews M, Bruford EA, Thompson DC, Vasiliou V, et al. Update on the human and mouse lipocalin (LCN) gene family, including evidence the mouse Mup cluster is result of an "evolutionary bloom". Hum Genomics. 2019;13(1):11.
2. Feyereisen R. Evolution of insect P450. Biochem Soc Trans. 2006;34(Pt 6):1252–5.
3. Nelson DR, Goldstone JV, Stegeman JJ. The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s. Philos Trans R Soc Lond B Biol Sci. 2013;368(1612):20120474.
4. Ohno S. Evolution by gene duplication. Springer; 1970. Available from: https://www.ncbi.nlm.nih.gov/nlmcatalog/?term=evolution%20by%20gene%20duplication%20ohno%201970.
5. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. The evolution of mammalian gene families. PLoS One. 2006;1:e85.
6. Giorgianni MW, Dowell NL, Griffin S, Kassner VA, Selegue JE, Carroll SB. The origin and diversification of a novel protein family in venomous snakes. Proc Natl Acad Sci U S A. 2020;117(20):10911–20.
7. Holding ML, Strickland JL, Rautsaw RM, Hofmann EP, Mason AJ, Hogan MP, et al. Phylogenetically diverse diets favor more complex venoms in North American pitvipers. Proc Natl Acad Sci U S A. 2021;118(17):e2015579118.
8. Johnson RN, O'Meally D, Chen Z, Etherington GJ, Ho SYW, Nash WJ, et al. Adaptation and conservation insights from the koala genome. Nat Genet. 2018;50(8):1102–11.
9. Pavlovich SS, Lovett SP, Koroleva G, Guito JC, Arnold CE, Nagle ER, et al. The Egyptian rousette genome reveals unexpected features of bat antiviral immunity. Cell. 2018;173(5):1098-1110.e18.
10. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39(10):1256–60.
11. Sun X, Zhang Z, Sun Y, Li J, Xu S, Yang G. Comparative genomics analyses of alpha-keratins reveal insights into evolutionary adaptation of marine mammals. Front Zool. 2017;14(1):41.
12. Kawasaki K, Lafont AG, Sire JY. The evolution of milk casein genes from tooth genes before the origin of mammals. Mol Biol Evol. 2011;28(7):2053–61.
13. Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, et al. Adaptive evolution of UGT2B17 copy-number variation. Am J Hum Genet. 2008;83(3):337–46.
14. Gaedigk A, Blum M, Gaedigk R, Eichelbaum M, Meyer UA. Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. Am J Hum Genet. 1991;48(5):943–50.
15. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol. 2010;11(3):R26.
16. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823–8.
17. Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A. 1997;94(15):7799–806.
18. Niimura Y, Nei M. Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. J Hum Genet. 2006;51(6):505–17.
19. Zhang J, Webb DM. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. Proc Natl Acad Sci U S A. 2003;100(14):8337–41.
20. Clowney EJ, Magklara A, Colquitt BM, Pathak N, Lane RP, Lomvardas S. High-throughput mapping of the promoters of the mouse olfactory receptor genes reveals a new type of mammalian promoter and

provides insight into olfactory receptor gene regulation. Genome Res. 2011;21(8):1249–59.
21. Corneo G, Ginelli E, Soave C, Bernardi G. Isolation and characterization of mouse and guinea pig satellite deoxyribonucleic acids. Biochemistry. 1968;7(12):4373–9.
22. Holmquist GP. Chromosome bands, their chromatin flavors, and their functional features. Am J Hum Genet. 1992;51(1):17–37.
23. Bickmore WA. Patterns in the genome. Heredity. 2019;123(1):50–7.
24. Filipski J. Evolution of DNA sequence contributions of mutational bias and selection to the origin of chromosomal compartments. In: Obe G, editor. Advances in mutagenesis research. Berlin, Heidelberg: Springer; 1990. p. 1–54. https://doi.org/10.1007/978-3-642-75599-6_1.
25. Korenberg JR, Rykowski MC. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. Cell. 1988;53(3):391–400.
26. Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ. Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. 2005;6(4):R33.
27. Waterson RH, Lander ES, Wilson RK, The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437(7055):69–87.
28. Niimura Y, Gojobori T. In silico chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. Proc Natl Acad Sci U S A. 2002;99(2):797–802.
29. Costantini M, Clay O, Auletta F, Bernardi G. An isochore map of human chromosomes. Genome Res. 2006;16(4):536–41.
30. Cohen N, Dagan T, Stone L, Graur D. GC composition of the human genome: in search of isochores. Mol Biol Evol. 2005;22(5):1260–72.
31. Cozzi P, Milanesi L, Bernardi G. Segmenting the human genome into isochores. Evol Bioinform Online. 2015;11:253–61.
32. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
33. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311.
34. Lynch M. The origins of genome architecture. Indiana University Press; 2007. Available from: https://www.ncbi.nlm.nih.gov/nlmcatalog/101296442.
35. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, et al. The mosaic genome of warm-blooded vertebrates. Science. 1985;228(4702):953–8.
36. Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. Quantification of GC-biased gene conversion in the human genome. Genome Res. 2015;25(8):1215–28.
37. Duret L, Arndt PF. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 2008;4(5):e1000071.
38. Pouyet F, Mouchiroud D, Duret L, Sémon M. Recombination, meiotic expression and human codon usage. Przeworski M, editor. eLife. 2017;6:e27344.
39. Fryxell KJ, Zuckerkandl E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol Biol Evol. 2000;17(9):1371–83.
40. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 2010;6(9):e1001115.
41. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010;6(9):e1001107.
42. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A. 2010;107(3):961–8.
43. Simmen MW. Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. Genomics. 2008;92(1):33–40.
44. Mugal CF, Arndt PF, Holm L, Ellegren H. Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes. G3 (Bethesda). 2015;5(3):441–7.
45. Fryxell KJ, Moon WJ. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol. 2005;22(3):650–8.
46. Jabbari K, Chakraborty M, Wiehe T. DNA sequence-dependent chromatin architecture and nuclear hubs formation. Sci Rep. 2019;9(1):14646.
47. Ramani V, Shendure J, Duan Z. Understanding spatial genome organization: methods and insights. Genomics Proteomics Bioinformatics. 2016;14(1):7–20.
48. Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, et al. Replication timing of the human genome. Hum Mol Genet. 2004;13(2):191–202.
49. Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet. 2014;15(1):47–70.
50. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A. 2006;103(5):1412–7.
51. Antequera F. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci. 2003;60(8):1647–58.
52. Armelin-Correa LM, Gutiyama LM, Brandt DYC, Malnic B. Nuclear compartmentalization of odorant receptor genes. Proc Natl Acad Sci U S A. 2014;111(7):2782–7.
53. Clowney EJ, LeGros MA, Mosley CP, Clowney FG, Markenskoff-Papadimitriou EC, Myllys M, et al. Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell. 2012;151(4):724–37.
54. Monahan K, Horta A, Lomvardas S. LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature. 2019;565(7740):448–53.
55. Tan L, Xing D, Daley N, Xie XS. Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. Nat Struct Mol Biol. 2019;26(4):297–307.
56. Tan L, Ma W, Wu H, Zheng Y, Xing D, Chen R, et al. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. Cell. 2021;184(3):741-758.e17.
57. Holmquist GP, Filipski J. Organization of mutations along the genome: a prime determinant of genome evolution. Trends Ecol Evol. 1994;9(2):65–9.
58. Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, et al. Evolutionary constraint and innovation across hundreds of placental mammals. bioRxiv; 2023. p. 2023.03.09.531574. Available from: https://www.biorxiv.org/content/10.1101/2023.03.09.531574v1.
59. Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res. 2010;20(1):1–9.
60. Liu A, He F, Shen L, Liu R, Wang Z, Zhou J. Convergent degeneration of olfactory receptor gene repertoires in marine mammals. BMC Genomics. 2019;20(1):977.
61. Li D, Zhang J. Diet shapes the evolution of the vertebrate bitter taste receptor gene repertoire. Mol Biol Evol. 2014;31(2):303–9.
62. Niimura Y, Nei M. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One. 2007;2(8):e708.
63. Jabbari K, Bernardi G. An isochore framework underlies chromatin architecture. PLoS One. 2017;12(1):e0168023.
64. Gao F, Zhang CT. GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. Nucleic Acids Res. 2006;34(Web Server issue):W686-91.
65. Clawson H. GC percent in 5-Base Windows (gc5BaseBw). In: UCSC Genome Browser. 2018. https://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=1677716454_0DkT8G7PCmkklWxCaNQv6aiHsfm8&db=hg38&c=chr21&g=gc5BaseBw. Accessed 17 Aug 2023.
66. Micklem G, Hillier LW. CpG Islands. In: UCSC Genomewiki. 2006. http://genomewiki.ucsc.edu/index.php/CpG_Islands. Accessed 17 Aug 2023.
67. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604(7905):310–5.
68. Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. Genome Res. 2001;11(5):685–702.
69. Niimura Y, Matsui A, Touhara K. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res. 2014;24(9):1485–96.
70. Khan I, Maldonado E, Vasconcelos V, O'Brien SJ, Johnson WE, Antunes A. Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments. BMC Genomics. 2014;15(1):779.

71. Walker MB, King BL, Paigen K. Clusters of ancestrally related genes that show paralogy in whole or in part are a major feature of the genomes of humans and other species. PLoS One. 2012;7(4):e35274.
72. Sheehan MJ, Campbell P, Miller CH. Evolutionary patterns of major urinary protein scent signals in house mice and relatives. Mol Ecol. 2019;28(15):3587–601.
73. Hardison RC. Evolution of hemoglobin and its genes. Cold Spring Harb Perspect Med. 2012;2(12):a011627.
74. Thomas JH. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 2007;3(5):e67.
75. Tan HM, Low WY. Rapid birth-death evolution and positive selection in detoxification-type glutathione S-transferases in mammals. PLoS One. 2018;13(12):e0209336.
76. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet. 2008;9(12):951–63.
77. Schwartz JC, Gibson MS, Heimeier D, Koren S, Phillippy AM, Bickhart DM, et al. The evolution of the natural killer complex; a comparison between mammals using new high-quality genome assemblies and targeted annotation. Immunogenetics. 2017;69(4):255–69.
78. Semple F, Dorin JR. β-Defensins: multifunctional modulators of infection, inflammation and more? J Innate Immun. 2012;4(4):337–48.
79. Shelton JF, Shastri AJ, Fletez-Brant K, Aslibekyan S, Auton A. The UGT2A1/UGT2A2 locus is associated with COVID-19-related loss of smell or taste. Nat Genet. 2022;54(2):121–4.
80. Chuang JH, Li H. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. PLoS Biol. 2004;2(2):E29.
81. Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, et al. The DNA sequence and biology of human chromosome 19. Nature. 2004;428(6982):529–35.
82. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
83. Mainland JD, Keller A, Li YR, Zhou T, Trimmer C, Snyder LL, et al. The missense of smell: functional variability in the human odorant receptor repertoire. Nat Neurosci. 2014;17(1):114–20.
84. Trimmer C, Keller A, Murphy NR, Snyder LL, Willer JR, Nagai MH, et al. Genetic variation across the human olfactory receptor repertoire alters odor perception. Proc Natl Acad Sci U S A. 2019;116(19):9475–80.
85. Rodriguez-Galindo M, Casillas S, Weghorn D, Barbadilla A. Germline de novo mutation rates on exons versus introns in humans. Nat Commun. 2020;11(1):3304.
86. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000;156(1):297–304.
87. Sved J, Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci U S A. 1990;87(12):4692–6.
88. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
89. Hughes GM, Boston ESM, Finarelli JA, Murphy WJ, Higgins DG, Teeling EC. The birth and death of olfactory receptor gene families in mammalian niche adaptation. Mol Biol Evol. 2018;35(6):1390–406.
90. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of *de novo* mutations in humans. Nature Genet. 2015;47(7):822–6.
91. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017;549(7673):519–22.
92. Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 2019;363(6425):eaau1043.
93. An JY, Lin K, Zhu L, Werling DM, Dong S, Brand H, et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science. 2018;362(6420):eaat6576.
94. Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, et al. Parent-of-origin-specific signatures of de novo mutations. Nat Genet. 2016;48(8):935–9.
95. Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20(4):602–11.
96. Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, et al. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Williams AL, McCarthy MI, Williams AL, editors. eLife. 2019;8:e46922.
97. Jabbari K, Wirtz J, Rauscher M, Wiehe T. A common genomic code for chromatin architecture and recombination landscape. PLoS One. 2019;14(3):e0213278.
98. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467(7319):1099–103.
99. Lee YS, Chao A, Chen CH, Chou T, Wang SYM, Wang TH. Analysis of human meiotic recombination events with a parent-sibling tracing approach. BMC Genomics. 2011;12(1):434.
100. Uhrberg M. The KIR gene family: life in the fast lane of evolution. Eur J Immunol. 2005;35(1):10–5.
101. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 2010;327(5967):836–40.
102. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science. 2010;327(5967):876–9.
103. Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. Science. 2010;327(5967):835.
104. Altemose N, Noor N, Bitoun E, Tumian A, Imbeault M, Chapman JR, et al. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. Przeworski M, editor. eLife. 2017;6:e28383.
105. Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, et al. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. de Massy B, editor. eLife. 2017;6:e24133.
106. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14(3):204–20.
107. Mohn F, Schübeler D. Genetics and epigenetics: stability and plasticity during cellular differentiation. Trends Genet. 2009;25(3):129–36.
108. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10):1010–22.
109. Ponger L, Duret L, Mouchiroud D. Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res. 2001;11(11):1854–60.
110. Elango N, Yi SV. Functional relevance of CpG island length for regulation of gene expression. Genetics. 2011;187(4):1077–83.
111. Michaloski JS, Galante PAF, Malnic B. Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. Genome Res. 2006;16(9):1091–8.
112. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321(6067):209–13.
113. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
114. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global epigenomic reconfiguration during mammalian brain development. Science. 2013;341(6146):1237905.
115. Winick-Ng W, Kukalev A, Harabula I, Zea-Redondo L, Szabó D, Meijer M, et al. Cell-type specialization is encoded by specific chromatin topologies. Nature. 2021;599(7886):684–91.
116. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. 2016;17(8):2042–59.
117. Naughton C, Avlonitis N, Corless S, Prendergast JG, Mati IK, Eijk PP, et al. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat Struct Mol Biol. 2013;20(3):387–95.
118. Witt M, Hummel T. Vomeronasal versus olfactory epithelium: is there a cellular basis for human vomeronasal perception? Int Rev Cytol. 2006:209–59. Available from: https://www.sciencedirect.com/science/article/pii/S0074769606480049.

119. Xia B, Yan Y, Baron M, Wagner F, Barkley D, Chiodin M, et al. Widespread transcriptional scanning in the testis modulates gene evolution rates. Cell. 2020;180(2):248-262.e21.
120. Harpak A, Lan X, Gao Z, Pritchard JK. Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. Proc Natl Acad Sci U S A. 2017;114(48):12779–84.
121. Caridi PC, Delabaere L, Zapotoczny G, Chiolo I. And yet, it moves: nuclear and chromatin dynamics of a heterochromatic double-strand break. Philos Trans R Soc Lond B Biol Sci. 2017;372(1731):20160291.
122. Jimeno S, Mejías-Navarro F, Prados-Carvajal R, Huertas P. Chapter Four - controlling the balance between chromosome break repair pathways. In: Donev R, editor. Adv Protein Chem Struct Biol. 2019:95–134. (DNA Repair; vol. 115). Available from: https://www.sciencedirect.com/science/article/pii/S187616231830066X.
123. Gu J, Lu H, Tsai AG, Schwarz K, Lieber MR. Single-stranded DNA ligation and XLF-stimulated incompatible DNA end ligation by the XRCC4-DNA ligase IV complex: influence of terminal DNA sequence. Nucleic Acids Res. 2007;35(17):5755–62.
124. Davis BJ, Havener JM, Ramsden DA. End-bridging is required for pol μ to efficiently promote repair of noncomplementary ends by nonhomologous end joining. Nucleic Acids Res. 2008;36(9):3085–94.
125. Grand RS, Burger L, Gräwe C, Michael AK, Isbel L, Hess D, et al. BANP opens chromatin and activates CpG-island-regulated genes. Nature. 2021:133–7.
126. Zhang J, Zhang Y, You Q, Huang C, Zhang T, Wang M, et al. Highly enriched BEND3 prevents the premature activation of bivalent genes during differentiation. Science. 2022;375(6584):1053–8.
127. Monahan K, Schieren I, Cheung J, Mumbey-Wafula A, Monuki ES, Lomvardas S. Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. Elife. 2017;6:e28620.
128. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. Locus control regions. Blood. 2002;100(9):3077–86.
129. Pratto F, Brick K, Cheng G, Lam KWG, Cloutier JM, Dahiya D, et al. Meiotic recombination mirrors patterns of germline replication in mice and humans. Cell. 2021;184(16):4251-4267.e20.
130. Xie WJ, Meng L, Liu S, Zhang L, Cai X, Gao YQ. Structural modeling of chromatin integrates genome features and reveals chromosome folding principle. Sci Rep. 2017;7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5460185/.
131. Costantini M, Musto H. The isochores as a fundamental level of genome structure and organization: a general overview. J Mol Evol. 2017;84(2):93–103.
132. Dekker J. GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol. 2007;8(6):R116.
133. Wang Z, Willard HF. Evidence for sequence biases associated with patterns of histone methylation. BMC Genomics. 2012;13(1):367.
134. Magklara A, Yen A, Colquitt BM, Clowney EJ, Allen W, Markenscoff-Papadimitriou E, et al. An epigenetic signature for monoallelic olfactory receptor expression. Cell. 2011;145(4):555–70.
135. Toyoda S, Kawaguchi M, Kobayashi T, Tarusawa E, Toyama T, Okano M, et al. Developmental epigenetic modification regulates stochastic expression of clustered protocadherin genes, generating single neuron diversity. Neuron. 2014;82(1):94–108.
136. Williams DL, Sikora VM, Hammer MA, Amin S, Brinjikji T, Brumley EK, et al. May the odds be ever in your favor: non-deterministic mechanisms diversifying cell surface molecule expression. Front Cell Dev Biol. 2021;9:720798.
137. Balan S, Iwayama Y, Ohnishi T, Fukuda M, Shirai A, Yamada A, et al. A loss-of-function variant in SUV39H2 identified in autism-spectrum disorder causes altered H3K9 trimethylation and dysregulation of protocadherin β-cluster genes in the developing brain. Mol Psychiatry. 2021;26(12):7550–9.
138. Nicetto D, Donahue G, Jain T, Peng T, Sidoli S, Sheng L, et al. H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification. Science. 2019;363(6424):294–7.
139. Antequera F, Bird A. CpG islands as genomic footprints of promoters that are associated with replication origins. Curr Biol. 1999;9(17):R661–7.
140. Horowitz A, Strauss-Albee DM, Leipold M, Kubo J, Nemat-Gorgani N, Dogan OC, et al. Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. Sci Transl Med. 2013;5(208):208ra145.
141. Wilk AJ, Blish CA. Diversification of human NK cells: lessons from deep profiling. J Leukoc Biol. 2018;103(4):629–41.
142. Dalton RP, Lyons DB, Lomvardas S. Co-opting the unfolded protein response to elicit olfactory receptor feedback. Cell. 2013;155(2):321–32.
143. Hetz C, Zhang K, Kaufman RJ. Mechanisms, regulation and functions of the unfolded protein response. Nat Rev Mol Cell Biol. 2020;21(8):421–38.
144. Guo Y, Monahan K, Wu H, Gertz J, Varley KE, Li W, et al. CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. Proc Natl Acad Sci USA. 2012;109(51):21081.
145. Canzio D, Nwakeze CL, Horta A, Rajkumar SM, Coffey EL, Duffy EE, et al. Antisense lncRNA transcription mediates DNA demethylation to drive stochastic protocadherin α promoter choice. Cell. 2019;177(3):639-653.e15.
146. Markenscoff-Papadimitriou E, Binyameen F, Whalen S, Price J, Lim K, Ypsilanti AR, et al. Autism risk gene POGZ promotes chromatin accessibility and expression of clustered synaptic genes. Cell Rep. 2021;37(10):110089.
147. Smukowski CS, Noor MAF. Recombination rate variation in closely related species. Heredity. 2011;107(6):496–508.
148. Jeffreys AJ, Cotton VE, Neumann R, Lam KWG. Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. Proc Natl Acad Sci U S A. 2013;110(2):600–5.
149. Casola C, Betrán E. The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? Genome Biol Evol. 2017;9(6):1351–73.
150. Lane RP, Young J, Newman T, Trask BJ. Species specificity in rodent pheromone receptor repertoires. Genome Res. 2004;14(4):603–8.
151. Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, Blankenship J, et al. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Hum Mol Genet. 1998;7(1):13–26.
152. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature. 2005;437(7055):94–100.
153. Mefford HC, Linardopoulou E, Coil D, van den Engh G, Trask BJ. Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. Hum Mol Genet. 2001;10(21):2363–72.
154. Newman T, Trask BJ. Complex evolution of 7E olfactory receptor genes in segmental duplications. Genome Res. 2003;13(5):781–93.
155. Kim J, Farré M, Auvil L, Capitanu B, Larkin DM, Ma J, et al. Reconstruction and evolutionary history of eutherian chromosomes. Proc Natl Acad Sci U S A. 2017;114(27):E5379–88.
156. Yue Y, Haaf T. 7E olfactory receptor gene clusters and evolutionary chromosome rearrangements. Cytogenet Genome Res. 2006;112(1–2):6–10.
157. Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ. Extensive copy-number variation of the human olfactory receptor gene family. Am J Hum Genet. 2008;83(2):228–42.
158. Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, et al. Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet. 2001;68(4):874–83.
159. Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, Giorda R, et al. Heterozygous submicroscopic inversions involving olfactory receptor–gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. Am J Hum Genet. 2002;71(2):276–85.
160. Maas NMC, Van Vooren S, Hannes F, Van Buggenhout G, Mysliwiec M, Moreau Y, et al. The t(4;8) is mediated by homologous recombination between olfactory receptor gene clusters, but other 4p16 translocations occur at random. J Genet Couns. 2007;18(4):357–65.
161. Paudel Y, Madsen O, Megens HJ, Frantz LAF, Bosse M, Crooijmans RPMA, et al. Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. BMC Genomics. 2015;16(1):330.
162. Rogers RL. Chromosomal rearrangements as barriers to genetic homogenization between archaic and modern humans. Mol Biol Evol. 2015;32(12):3064–78.
163. North HL, Caminade P, Severac D, Belkhir K, Smadja CM. The role of copy-number variation in the reinforcement of sexual isolation

between the two European subspecies of the house mouse. Philos Trans R Soc Lond B Biol Sci. 1806;2020(375):20190540.
164. Otto M, Zheng Y, Wiehe T. Recombination, selection, and the evolution of tandem gene arrays. Genetics. 2022;221(3):iyac052.
165. Qu Q, Haitina T, Zhu M, Ahlberg PE. New genomic and fossil data illuminate the origin of enamel. Nature. 2015;526(7571):108–11.
166. Mann RS. Why are Hox genes clustered? BioEssays. 1997;19(8):661–4.
167. Darbellay F, Bochaton C, Lopez-Delisle L, Mascrez B, Tschopp P, Delpretti S, et al. The constrained architecture of mammalian Hox gene clusters. Proc Natl Acad Sci U S A. 2019;116(27):13424–33.
168. Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, Splinter E, et al. A regulatory archipelago controls Hox genes transcription in digits. Cell. 2011;147(5):1132–45.
169. Wong ES, Zheng D, Tan SZ, Bower NL, Garside V, Vanwalleghem G, et al. Deep conservation of the enhancer regulatory code in animals. Science. 2020;370(6517):eaax8137.
170. Markenscoff-Papadimitriou E, Allen WE, Colquitt BM, Goh T, Murphy KK, Monahan K, et al. Enhancer interaction networks as a means for singular olfactory receptor expression. Cell. 2014;159(3):543–57.
171. Roy AL, Sen R, Roeder RG. Enhancer–promoter communication and transcriptional regulation of Igh. Trends Immunol. 2011;32(11):532–9.
172. Ribich S, Tasic B, Maniatis T. Identification of long-range regulatory elements in the protocadherin-α gene cluster. Proc Natl Acad Sci U S A. 2006;103(52):19719–24.
173. Yokota S, Hirayama T, Hirano K, Kaneko R, Toyoda S, Kawamura Y, et al. Identification of the cluster control region for the protocadherin-beta genes located beyond the protocadherin-gamma cluster. J Biol Chem. 2011;286(36):31885–95.
174. de Lange T. A loopy view of telomere evolution. Front Genet. 2015;6. Available from: https://www.frontiersin.org/article/10.3389/fgene.2015.00321.
175. Zody MC, Garber M, Adams DJ, Sharpe T, Harrow J, Lupski JR, et al. DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature. 2006;440(7087):1045–9.
176. Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu XL, Mudunuri U, et al. Mapping and initial analysis of human subtelomeric sequence assemblies. Genome Res. 2004;14(1):18–28.
177. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, et al. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 2004;14(4):528–38.
178. Hultén M. Chiasma distribution at diakinesis in the normal human male. Hereditas. 1974;76(1):55–78.
179. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. Recombination initiation maps of individual human genomes. Science. 2014;346(6211):1256442.
180. Tease C, Hultén MA. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. Cytogenet Genome Res. 2004;107(3–4):208–15.
181. Antunes DMF, Kalmbach KH, Wang F, Dracxler RC, Seth-Smith ML, Kramer Y, et al. A single-cell assay for telomere DNA content shows increasing telomere length heterogeneity, as well as increasing mean telomere length in human spermatozoa with advancing age. J Assist Reprod Genet. 2015;32(11):1685–90.
182. Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. PLoS Genet. 2013;9(8):e1003684.
183. Huttener R, Thorrez L, in't Veld T, Granvik M, Snoeck L, Van Lommel L, et al. GC content of vertebrate exome landscapes reveal areas of accelerated protein evolution. BMC Evol Biol. 2019;19(1):144.
184. Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, et al. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature. 2022;602(7895):101–5.
185. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009;5(5):e1000471.
186. Murphy D, Elyashiv E, Amster G, Sella G. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. bioRxiv; 2021. p. 2021.07.02.450762. Available from: https://www.biorxiv.org/content/10.1101/2021.07.02.450762v2.
187. Kambere MB, Lane RP. Exceptional LINE density at V1R loci: the Lyon repeat hypothesis revisited on autosomes. J Mol Evol. 2009;68(2):145–59.
188. Fraimovitch E, Hagai T. Promoter evolution of mammalian gene duplicates. BMC Biol. 2023;21(1):80.
189. Lan X, Pritchard JK. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science. 2016;352(6288):1009–13.
190. Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. Nat Rev Genet. 2019;20(8):437–55.
191. McKenzie SK, Fetter-Pruneda I, Ruta V, Kronauer DJC. Transcriptomics and neuroanatomy of the clonal raider ant implicate an expanded clade of odorant receptors in chemical communication. Proc Natl Acad Sci USA. 2016;113(49):14091–6.
192. Armitage SAO, Freiburg RY, Kurtz J, Bravo IG. The evolution of Dscam genes across the arthropods. BMC Evol Biol. 2012;12:53.
193. Labrador M, Corces VG. Extensive exon reshuffling over evolutionary time coupled to trans-splicing in Drosophila. Genome Res. 2003;13(10):2220–8.
194. Goeke S, Greene EA, Grant PK, Gates MA, Crowner D, Aigaki T, et al. Alternative splicing of lola generates 19 transcription factors controlling axon guidance in Drosophila. Nat Neurosci. 2003;6(9):917–24.
195. Venables JP, Tazi J, Juge F. Regulated functional alternative splicing in Drosophila. Nucleic Acids Res. 2012;40(1):1–10.
196. Schield DR, Card DC, Hales NR, Perry BW, Pasquesi GM, Blackmon H, et al. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. Genome Res. 2019;29(4):590–601.
197. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome Trypanosoma brucei. Science. 2005;309(5733):416–22.
198. Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, et al. caper: comparative analyses of phylogenetics and evolution in R. 2018. Available from: https://cran.r-project.org/web/packages/caper/index.html.
199. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 2017;34(7):1812–9.
200. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526–8.
201. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
202. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44(Web Server issue):W160-5.
203. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innov J. 2021;2(3):100141.
204. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25(2):288–9.
205. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res. 2021;49(D1):D939-46.
206. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184–91.
207. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.
208. Kim J, Lee C, Ko BJ, Yoo DA, Won S, Phillippy AM, et al. False gene and chromosome losses in genome assemblies caused by GC content variation and repeats. Genome Biol. 2022;23(1):204.
209. Scripts for tandemClipR. Available from: https://github.com/Matthew-Holding/TandemClipR. Accessed 17 Aug 2023.
210. Track Hub. Available from: http://genome.ucsc.edu/s/mbrovkin/hg38. Accessed 17 Aug 2023.

Publisher's Note